

High-throughput modelling and structural investigation of cysteine protease complexes with protein inhibitors

A thesis submitted in fulfillment of the requirement for the degree

MASTER OF SCIENCE

in

BIOCHEMISTRY

of

RHODES UNIVERSITY

by

Matthys Christoffel Kroon

November 18, 2012

Abstract

The papain-like cysteine protease family (C1 proteases) is highly important because of its involvement in research and industrial applications and its role in various human diseases. Protein inhibitors are an important aspect of C1 protease biology and are relevant to its clinical, industrial and research importance. To study the interaction between the proteases and the inhibitors it is very useful to have accurate structural models of the protease-inhibitor complexes. To this end, a high-throughput pipeline for modelling complexes of papain-like cysteine proteases and protein inhibitors was implemented and tested (Tastan Bishop & Kroon, 2011). The pipeline utilizes a novel technique for obtaining modelling templates by using superpositioning to combine coordinates from separate experimental structures. To test the pipeline, models of complexes with known structures (test set) were modelled using many different templates and the resultant models evaluated to compare the quality of the different templates. It was found that use of the new technique to obtain templates did not introduce significant errors, while allowing closer homologs to be used for modelling - leading to more accurate models. The test set models were also used to evaluate certain steps of the modelling protocol. The effect of Rosetta energy minimization on model accuracy and the use of Rosetta energy and DOPE Z-score values to identify accurate models were investigated. Several complexes were then modelled using the best available templates according to criteria informed by the previous results. A website was built that allows a user to download any of the metrics or models produced in the study. This website is accessible at <http://rubi.ru.ac.za/cpmdb>.

Declaration

I, Matthys Christoffel Kroon, hereby declare that the work presented here is my own, original, work and has not been submitted for examination for any degree at any other educational institution.

03/06/2012 _____

Acknowledgements

I'd like to express my sincere gratitude to all the housemates, labmates, friends and family that supported me during the last three years. I am very lucky and privileged to have you. I'd also like to thank the senior bioinformatics students at the University of Pretoria who provided valuable guidance, especially in 2008. I'd like to specifically mention Pieter Burger and Charles Hefer who taught me a great deal. I'd like to thank my parents for affording me the opportunity to pursue further studies. Without their help, it would not have been possible. To my co-supervisor, Prof. Fourie Joubert, a big thank you for infrastructure, moral and academic support. To my supervisor, Dr Özlem Tastan Bishop, thanks for providing guidance, infrastructure and scientific input. Your dedication to your students is inspiring. Lastly, I'd like to thank Rhodes University and the National Bioinformatics Network for providing the funding which made this project possible.

Research output

Article:

Tastan Bishop, Özlem, & Kroon, Matthys. 2011. Study of protein complexes via homology modeling, applied to cysteine proteases and their protein inhibitors. *Journal of Molecular Modeling*. **17**(12), 3163-72.

Posters:

Özlem Tastan Bishop and Matthys Kroon, “An alternative approach to protein docking: Complexes via homology modeling” 19th Annual International Conference on Intelligent Systems for Molecular Biology & 10th European Conference on Computational Biology (ISMB-ECCB), Vienna, Austria, July 2011.

Tastan Bishop, Ö. and Kroon, M. “A Homology Based Study of the Complexes of Cysteine Proteases and Their Protein Inhibitors” ISCB Africa ASBCB Conference on Bioinformatics, Cape Town, South Africa, 9-11 March 2011.

Kroon, M., Joubert, F., and Tastan Bishop, Ö. “Large Scale Sequence and Structural Analysis of Papain-like Cysteine Proteases” SAICSIT WITS conference 2009, 11-12 October 2009. (<http://www.bioinf.wits.ac.za/sabioinf2009/biopapers.html>)

Tastan Bishop, A.Ö. and Kroon, M. “Identification and Analysis of Common Features of Cysteine Proteases of Malaria Parasites for the Development of Novel Parasitic Protease Inhibitors” 17th Annual International Conference on Intelligent Systems for Molecular Biology & 8th European Conference on Computational Biology (ISMB-ECCB), Stockholm, Sweden. June 2009.

Contents

1	Introduction	12
1.1	Chapter overview	12
1.2	Malaria	12
1.3	Papain-like cystein proteases	13
1.3.1	Summary	13
1.3.2	History	13
1.3.3	Classification	13
1.3.4	Structure and function	14
1.3.5	Biological roles	15
1.3.6	Regulation	15
1.3.7	Clinical significance	16
1.3.8	Economical significance	17
1.3.9	Cysteine proteases of malaria parasites	17
1.3.10	Available data	18
1.4	Cystatins	19
1.4.1	Summary	19
1.4.2	History	19
1.4.3	Classification	19
1.4.4	Structure and function	20
1.4.5	Biological roles	21
1.4.6	Clinical significance	21
1.4.7	Economical significance	22
1.4.8	Available data	22

1.5	Chagasin-like inhibitors	22
1.5.1	Summary	22
1.5.2	History	23
1.5.3	Classification	23
1.5.4	Structure and function	23
1.5.5	Biological roles	23
1.5.6	Clinical significance	24
1.5.7	Economical significance	24
1.5.8	Chagasin-like inhibitors of malarial proteases	25
1.5.9	Available data	25
1.6	Comparison between cystatin and chagasin	25
1.7	Computer-based protein structure prediction	26
1.8	Homology modelling	28
1.8.1	Introduction	28
1.8.2	Process overview	29
1.8.3	Modeller	31
1.9	Protein-protein interactions	32
1.9.1	Docking	33
1.9.2	Homology modelling of protein complexes	34
1.10	Data and databases	34
1.10.1	UniProt (www.uniprot.org)	34
1.10.2	Protein Data Bank (PDB, www.pdb.org)	35
1.10.3	MEROPS (merops.sanger.ac.uk)	35
1.10.4	Pfam (pfam.sanger.ac.uk)	35
1.11	Summary	35
2	Project design and rationale	36
2.1	Chapter overview	36
2.2	Project aim and motivation	36
2.3	Experimental design	38
2.3.1	Part 1: Examination of experimental structures	38

2.3.2	Part 2: Construction and evaluation of test set models	38
2.3.3	Part 3: Construction and evaluation of study set models	39
2.3.4	Part 4: Publication of results in website	41
2.4	Modelling overview	41
2.4.1	Data retrieval	41
2.4.2	Sequence alignment	41
2.4.3	Generation of templates	43
2.4.4	Template selection	43
2.4.5	Modelling protocol	46
2.4.6	Calculation of scoring functions	46
2.4.7	Energy minimization	46
2.5	Information management	46
2.6	Computational resources	47
3	Examination of crystal structures	48
3.1	Chapter overview	48
3.2	Aim and motivation	48
3.3	Method	49
3.4	Results and discussion: Papain-like cysteine proteases	51
3.4.1	Comparison between cathepsin-L like and cathepsin-B like subfamilies	51
3.4.2	Cathepsin-L like subfamily	51
3.4.3	Cathepsin-B like subfamily	54
3.5	Results and discussion: Cystatin superfamily	57
3.5.1	Comparison between type I and type II cystatins	57
3.5.2	Type I cystatins	59
3.5.3	Type II cystatins	61
3.6	Results and discussion: Chagasin-like inhibitors	62
3.7	Conclusion	66

4	Construction and evaluation of test models	67
4.1	Chapter overview	67
4.2	Aim and motivation	67
4.3	Method	68
4.3.1	Targets selected	68
4.3.2	Models built	68
4.3.3	Calculation of scoring functions	69
4.3.4	Evaluation metrics calculated	69
4.3.5	Energy minimization carried out	70
4.4	Results and discussion	71
4.4.1	General model accuracy	71
4.4.2	Direct comparison of simple and hybrid templates	76
4.4.3	Examination of relation between model accuracy and sequence identity and coverage	78
4.4.4	Examination of relationship between scoring functions and model accuracy	83
4.4.5	Examination of minimization results	86
4.4.5.1	Effect on accuracy	86
4.4.5.2	Are the energy scores more accurate for the minimized models?	90
4.4.5.3	Examination of relation between the energy change during minimization with the accuracy of the model	90
4.5	Conclusion	92
5	Construction of novel models	94
5.1	Chapter overview	94
5.2	Aim and motivation	94
5.3	Method	95
5.3.1	Targets selected	95
5.3.2	Template selection	95
5.3.3	Model building, energy minimization and calculation of scoring functions	95
5.4	Results and discussion	95

5.4.1	Template selection	95
5.4.2	Scoring function results before minimization	98
5.4.3	Scoring function results after minimization	98
5.5	Conclusion	98
6	Creation of website	101
6.1	Chapter overview	101
6.2	Aim and motivation	101
6.3	Method	101
6.4	Results and discussion	102
6.4.1	General	102
6.4.2	Alignment section	104
6.4.3	Download section	104
6.4.4	Browsing section	104
6.5	Conclusion	105
7	Conclusion	106
7.1	Chapter Overview	106
7.2	Conformational change can occur during complex formation between C1 proteases and protein inhibitors	107
7.3	Model accuracy could be estimated sensibly by measuring similarity to ex- perimental structures	107
7.4	There was no significant difference between the accuracy of hybrid and sim- ple models	108
7.5	Scoring functions provide useful information in certain circumstances	108
7.6	Rosetta energy minimization improves the accuracy of some models	109
7.7	Models for novel complexes were built and made available in a website . . .	109
7.8	Closing	110
	Bibliography	111

List of Figures

1.1	Structure of papain.	14
1.2	Structure of cystatin-B.	21
1.3	Structure of chagasin.	24
1.4	Comparison of cystatin-B and chagasin structures.	25
1.5	Homology modelling overview flowchart.	29
2.1	Homology modelling flowchart for study.	42
2.2	Complete template set.	44
2.3	Cystatin containing and chagasin containing templates.	45
2.4	Total numbers of models built.	45
3.1	Cathepsin-L and cathepsin-B structural	52
3.2	Global comparison of papain structures.	52
3.3	Interface comparison of papain structures.	53
3.4	Global comparison of cathepsin-B conformations in 1GMY.	54
3.5	Global comparison of cathepsin-B structures.	56
3.6	Interface comparison of cathepsin-B structures.	56
3.7	Global comparison of type I and type II cystatins.	58
3.8	Global comparison of cystatin-A structures.	59
3.9	Interface comparison of cystatin-A structures.	60
3.10	Global comparison of ovocystatin structures.	61
3.11	Interface comparison of ovocystatin structures.	62
3.12	Comparison of ovocystatin crystal structure and NMR solution structures. . .	63
3.13	Global comparison of chagasin structures.	64
3.14	Interface comparison of chagasin structures.	65

4.1	Distribution of major evaluation metrics for all test set models.	75
4.2	Standard deviation of RMSD values within a modelgroup as a function of the mean RMSD of that modelgroup.	76
4.3	Simple to hybrid template comparison set.	77
4.4	Comparison of evaluation metrics between simple and hybrid templates in representative subset.	79
4.5	RMSD as a function of sequence identity and sequence coverage in the test set.	81
4.6	i-RMSD as function of sequence identity and sequence similarity in the test set.	82
4.7	Rosetta energy as a function of RMSD in the test set.	84
4.8	DOPE Z-score as a function of RMSD in the test set.	85
4.9	Correlation between scoring functions and accuracy on the modelgroup level in the test set.	87
4.10	Correlation between scoring functions and RMSD on the modelgroup level as a function of RMSD in the test set.	88
4.11	Changes in evaluation metrics due to minimization in the test set.	89
4.12	Correlation between scoring functions and RMSD on the modelgroup level as a function of RMSD for the energy-minimized test set models.	91
4.13	Changes in scoring functions due to energy minimization in the test set. . . .	92
6.1	Screenshot of the website homepage	103

List of Tables

2.1	Experimental structures used in test set.	39
2.2	Study set proteins.	40
3.1	Structures used for investigation of solved structures.	50
3.2	Backbone RMSD in Å between selected structures of papain.	53
3.3	Backbone RMSD in Å between different cathepsin-B conformations of 1GMV.	55
3.4	Backbone RMSD in Å between selected structures of cathepsin-B.	57
3.5	Backbone RMSD in Å values between indicated structures of cystatin-A.	60
3.6	Backbone RMSD in Å between NMR solution structures of 1A67 and their averaged structure.	63
3.7	RMSD values of indicated chagasin structures.	65
4.1	Median test set RMSD results for cystatin-containing targets.	72
4.2	Median test set RMSD results for chagasin-containing targets.	73
5.1	Study set proteins	96
5.2	Templates selected for study set proteins with sequence identity, coverage and protein size.	97
5.3	Median scoring function results for study set models.	99
5.4	Median scoring function results for study set models after minimization.	100

Chapter 1

Introduction

1.1 Chapter overview

The purpose of this chapter is to familiarize the reader with some of the information and concepts relevant to the project. Because this project was originally inspired by work on cysteine proteases from malaria parasites, and contains some further work on these enzymes, the chapter starts off with a very brief discussion of malaria. It goes on by introducing the various protein families involved in the project. The papain-like cysteine proteases are introduced first, followed by the cystatins and the chagasin-like inhibitors. For each of these families, it gives some information regarding the history, function, classification and importance of the family. Computer based structure prediction is then discussed briefly. Homology modelling and protein docking as structure prediction techniques are introduced and explained shortly. Thereafter, the main sources of data used in the project are introduced and some information about each is given. The chapter ends with a very brief summary.

1.2 Malaria

Malaria is a parasitic disease that causes fever and other flu-like symptoms. It is a lethal disease, resulting in an estimated 767 000 deaths in 2008 (WHO, 2009). Young children and pregnant women are especially vulnerable. There are two species responsible for almost all cases of clinical malaria, they are *Plasmodium falciparum* and *Plasmodium vivax*. *P. falciparum* is by far the more lethal of the two species but *P. vivax* can cause long term infection with recurring symptoms (Price *et al.* , 2007). Drug resistance is a major problem in the fight against malaria (Hayton & Su, 2004). There is currently an international program for the eradication of malaria and a discontinuity in the ability to treat the disease effectively, using drugs, would be a major setback (WHO, 2009). Thus, there is a strong need for novel anti-malarial drugs. Cysteine proteases, in particular falcipain-2 of *P. falciparum*, have been well characterized and are considered very promising drug targets. There are a number of

potential falcipain-2 inhibitors currently under investigation as drugs for malaria (Ettari *et al.*, 2010).

1.3 Papain-like cystein proteases

1.3.1 Summary

Papain-like cysteine proteases (C1 proteases) are a very important and well-studied enzyme family. They catalyze the hydrolytic cleavage of peptide bonds. Their most common functions are in digestion and protein turnover but there are several other, more specialized, functions. They have been found to play roles in many parasitic processes and non-pathogenic human diseases. Several human and parasite C1 proteases have been proposed as drug targets and some inhibitors are currently in clinical trials. Protein inhibitors of C1 proteases are also of interest in agricultural biotechnology.

1.3.2 History

The C1 protease family is one of the oldest known enzyme families. The first member of the C1 family to be isolated and studied was papain, a cysteine protease that occurs naturally in the latex of the papaya fruit (*Carica papaya*). The name papain derives from the name of the fruit - papaya. In a 1885 paper which reported the isolation and analysis of papain it is also mentioned that the leaves and juice of papaya fruit was used to tenderize meat in the countries where it naturally occurred (Martin, 1885). In 1962, the three-dimensional (3D) structure of papain was one of the earliest protein structures to be published (Drenth *et al.*, 1962). Since the discovery of papain, a great many more C1 proteases, occurring in every living kingdom, have been identified and studied.

1.3.3 Classification

The classification of proteases and their inhibitors described here is the classification schema implemented by the MEROPS database (Rawlings *et al.*, 2010). A list of the clans and families in the database can be viewed at the website <http://merops.sanger.ac.uk>

Cysteine proteases as a group are defined as enzymes that catalyze the hydrolytic cleavage of peptide bonds, utilizing a catalytic cystein residue. They are divided into clans which are further divided into families. Cysteine proteases are grouped together in a clan if they are thought to share an evolutionary common ancestor. The different clans are thus thought to be products of independent evolution. Clan-members are grouped together in a family if there is detectable sequence similarity between them. The classification as a cysteine protease is thus dependent on a specific function while the subdivision are dependent on evolutionary history.

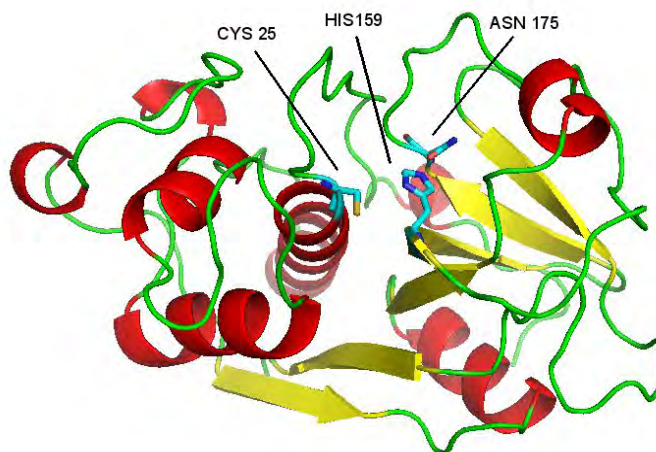


Figure 1.1 – The molecular structure of papain (PDB ID: 1STF) in cartoon representation. The L-domain is on the left and the R-domain on the right. The residues of the catalytic triad are shown in stick representation and their atoms are colored according to elements. The asparagine is visible on top, cysteine in the left and histidine in the middle.

The clans are identified by a single capital alphabetic letter ie. A, B, C. This is appended to the letter C (for Cysteine) to give a clan identifier eg. CA, CB and CC that uniquely identifies the protease clan. As homology between different clans, previously thought unrelated, became apparent, many of the clans were merged and thus the letters do not follow the alphabetic order. Families are identified by a numeric number. The number is unique across all the clans and not just within one clan. Thus, we only get one family C1, C2, etc. and not one in every clan. It is therefore redundant to specify the clan when specifying the family.

The papain-like cysteine protease family has received the family identifier C1 and falls in clan CA. It is the largest family of cysteine proteases and is almost universally distributed throughout the living kingdoms (Rawlings & Barrett, 1994). Within the C1 family there are 2 large sub-families namely the cathepsin-L like and cathepsin-B like subfamilies. The differentiating characteristic between them is the presence of a short insert, about 20 amino acids long, close to the active site in the cathepsin-B like subfamily. This insert forms a rather flexible loop that can interact with the substrate and with inhibitors (Sajid & McKerrow, 2002; Redzynia *et al.* , 2008). There are also other sub-families in the C1 family but they have been left out of this project due to insufficient data.

1.3.4 Structure and function

The enzyme consists of 2 distinct domains named the L and R domains. The catalytic cysteine residue is part of an α -helix in the L domain. The active site of the enzyme is at the interface of the 2 domains where a cleft is formed. The cleft contains binding pockets so that a polypeptide can bind to the cleft in an elongated conformation with the side-chains of its

amino acids accommodated in the binding pockets. This ensures the correct orientation of the labile bond to the catalytic cysteine residue.

The catalytic mechanism of cysteine proteases is similar to that of serine proteases, involving nucleophilic attack. The thiol group of cysteine is a better nucleophile than the hydroxyl group of serine because of the extra electron shell on the sulphur atom. The thiol group's nucleophilic character is increased by a nearby histidine that acts as a base and accepts the proton of the thiol group to yield a very nucleophilic thiolate group. The histidine then carries an imidazolium group. The two groups are closely associated in space and together form a charge relay diad. A conserved nearby asparagine residue serves as the oxyanion hole (Vernet *et al.* , 1995). Cysteine proteases require a reducing environment to function and this condition is usually met under cellular conditions by the presence of glutathione (Sajid & McKerrow, 2002). The structure of papain is shown in Figure 1.1. Using papain numbering, the catalytic residues are CYS 25, HIS 159 and ASN 175.

1.3.5 Biological roles

The human members of the family are known as the cathepsins and have important roles in the immune system, the protein renewal process and resorption of bone and cartilage. An important factor of their significance is that certain cathepsins are the only enzymes capable of degrading key proteins of the extracellular matrix. Eleven cathepsins have been described and are identified by letters of the alphabet. The letters B, C, F, H, K, L, O, S, V, W and X have been assigned. The biological roles they play have been reviewed by Vasiljeva *et al.* (2007).

C1 proteases of parasitic organisms play diverse and often important roles in parasite's life-cycles. The most common function is digestion. Roles in encystment, immune evasion, cell invasion and others are also important. An extensive review was compiled by Sajid & McKerrow (2002).

1.3.6 Regulation

Protein degradation is an activity that needs to be tightly controlled to prevent the harmful degradation of the wrong proteins. This regulation is carried out using three main mechanisms: synthesis as an inactive precursor, subcellular containment and inhibition by protein inhibitors (Sajid & McKerrow, 2002; Turk *et al.* , 2002).

C1 proteases are synthesized with an N-terminal prodomain that blocks access to the active site cleft. This prodomain has to be cleaved off for the enzyme to become active. The prodomain is also required for the correct folding of the enzyme although some C1 members from malaria parasites have been found to fold correctly in the absence of the prodomain. The cleavage of the prodomain and subsequent activation of the enzyme is often triggered by changes in pH. Sometimes the prodomain is cleaved off by other proteases when acting in protease cascades.

C1 proteases are usually localized to subcellular structures or excreted from the cell. The most common subcellular structures are lysosomes and digestive vacuoles. This prevents the indiscriminate degradation of useful proteins. Subcellularly-contained proteases are often released upon cell damage or necrosis and this can cause problems for the organism.

C1 activity is also regulated by protein inhibitors. The cystatin superfamily of inhibitors is the most common and most well studied family of protein inhibitor. Another, more recently discovered, family is the chagasin-like inhibitors. More details about the inhibitors are given in the sections dedicated to them (Section 1.4 and 1.5).

1.3.7 Clinical significance

Both human and non-human-derived C1 proteases play roles in disease and have been identified as promising drug targets. A lot of progress has been made in developing cysteine protease inhibitors for use as drugs (McKerrow *et al.* , 2008).

C1 proteases from parasitic organisms are involved in many diseases. Functions include digestion of proteins for nutrition, ex/en-cystment, cell/tissue invasion and immuno-modulation. These proteases are also usually very immunogenic. C1 proteases have been identified as important in various parasites for example:

- Malaria - Digestion of host hemoglobin for nutrients, erythrocyte invasion and erythrocyte rupture (Rosenthal, 2004).
- Liver flukes - Tissue invasion, immuno modulation and digestion (Alcala-Canto *et al.* , 2007; Dowd *et al.* , 1994; Dalton *et al.* , 2003).
- Amoebiasis - Tissue invasion (Freitas *et al.* , 2009; Tillack *et al.* , 2006; Moncada *et al.* , 2003).
- Hookworm - Digestion of host hemoglobin for nutrients (Ranjit *et al.* , 2009; Williamson *et al.* , 2004).
- Bilharzia - Digestion of host hemoglobin for nutrients (Caffrey *et al.* , 2002; Dvorák *et al.* , 2009).
- Leishmaniasis - Unclear (Mahmoudzadeh-Niknam & McKerrow, 2004; Mottram *et al.* , 2004).
- Chagas disease - Function unclear (Cazzulo *et al.* , 1997).

Human C1 proteases (cathepsins) are also involved in disease. A good review of the functions and disease involvement of human C1 proteases is Vasiljeva *et al.* (2007). For example, cathepsins can be involved in the following diseases:

- Cancer - Cathepsin-B and cathepsin-L play a role in tissue invasion by degrading the extracellular matrix, creating space for the tumor to grow. They are also involved in regulating cell-proliferation and apoptosis.
- Immune system and viral infections - Cathepsins are involved in processing antigen to be displayed on MHCII and also the process where the antigen fragment is loaded onto the MHCII molecule. Some viruses only become active after cleavage by cathepsins.
- Osteoporosis - Cathepsin-K plays a major role in bone resorption and degradation of the bone matrix. Cathepsin-L also plays a role as indicated by knock-out mice although the form of the involvement is still unknown. Osteoporosis is an imbalance in the normal bone resorption process, leading to weakened bones.
- Rheumatoid Arthritis - Cathepsin-B, -L and -K are thought to be involved in the destruction of bone tissue in this auto-immune disease.

With such a wide array of diseases it is clear why there is a lot of research interest in C1 proteases. If a library of non-toxic inhibitors were available with known inhibition profiles for all the cathepsins and important parasitic proteases the number of diseases that could be treated is large.

1.3.8 Economical significance

The clinical importance of C1 proteases also imply an economical importance as diseases cause loss of productivity and cost money to treat. In addition, C1 proteases have various industrial and research applications of economical significance. For example papain is used as meat-tenderizer (Ashie *et al.*, 2002). In many plants, C1 proteases act as a defense against herbivorous insects and in turn the insects use C1 proteases to digest plant material (Shindo & Van der Hoorn, 2008). This is important in agriculture. Papain is sometimes used in the cosmetic industry as a skin or hair softener although it can lead to allergic reactions (Sim *et al.*, 2000; Soto-Mera *et al.*, 2000). In biochemical research, papain is used for the cleavage of antibodies into variable and non-variable fragments. This allows researchers to carry out experiments that would not be possible otherwise. Medical applications of C1 proteases include the treatment of dental caries (Lopes *et al.*, 2007). The use of papain in topical treatments of chronic wounds is being discontinued due to side-effects, mainly allergies.

1.3.9 Cysteine proteases of malaria parasites

The falcipains are a group of 4 related cysteine proteases of the *P. falciparum* species. The group is made up out of falcipain-1, falcipain-2, falcipain-2' and falcipain-3. For an excellent introduction to the falcipains, see Ettari *et al.* 2010. The sequencing of the *P. falciparum* genome has greatly aided the identification and study of these enzymes. The falcipains have

been extensively studied and characterized largely because they have been identified as drug targets to treat malaria (Rosenthal, 2004).

Various studies have been carried out to determine the function of each of the falcipains. The best characterized function is hemoglobin degradation by falcipain-2 and falcipain-3 (Shenai *et al.* , 2000; Sijwali *et al.* , 2001). Treatment of malaria with cysteine protease inhibitors prevents growth of cultures *in vitro*, and cures malaria in mouse models (Rosenthal *et al.* , 1993). Another function under investigation is degradation of erythrocyte cytoskeletal proteins (Hanspal *et al.* , 2002). This is necessary for erythrocyte rupture during the parasite life-cycle. The main candidate for cytoskeletal degradation is falcipain-2, due to pH considerations. Falcipain-1 does not appear to be essential to the organism, as strains, in which the gene has been blocked, develops normally *in vitro*. However, falcipain-1 may play a role in functions that are not required *in vitro* such as immuno-evasion or other interactions with the host or vector. Falcipain-2' was discovered last and is almost identical to falcipain-2, showing over 90% sequence identity. There is a slight difference between the affinities of falcipain-2 and falcipain-2' for certain cytoskeletal proteins (Jeong *et al.* , 2006).

The structure of falcipain-2 was experimentally determined using X-ray diffraction in 2006 (Wang *et al.* , 2006). This indicated two novel features not found in other cysteine proteases of the papain family. The two features have been named the "arm" and "nose". The nose is thought to enable the protease to fold properly in the absence of the prodomain, a feat unique to falcipain-2 and related *Plasmodium* cysteine proteases. Residues have been identified that are important for the functioning of the "nose" motif via mutagenic experiments. The "arm" motif is thought to bind to hemoglobin. Deletion of the "arm" feature leads to loss of function against hemoglobin, while activity against artificial peptide substrates are unaffected.

The vivapains are a group of cysteine proteases in *P. vivax*. The vivapains are orthologs of the falcipains and share significant sequence similarity. A lot of what is known about the vivapains are related to studies of the falcipains, as *P. vivax* cannot be readily cultured like *P. falciparum*. At least one biochemical characterization of vivapain-2 and vivapain-3 has been carried out (Na *et al.* , 2004b). This showed slight inconsistencies with the characteristics of the falcipains, but both of the enzymes are able to degrade native hemoglobin as well as cytoskeletal proteins. Thus, while the specifics may not be exactly like that of the falcipains, vivapain-2 and 3 are thought to be involved in the same broad processes as falcipain-2 and 3. The vivapains are also potential drug targets and some hope that a general inhibitor may be designed to inhibit both the falcipains and the vivapains. An analysis of the variation in vivapain sequences within different *P. vivax* populations have been carried out and support their potential as drug targets (Na *et al.* , 2004a).

1.3.10 Available data

As a result of the importance of the C1 family, and the long time it's been under study, there is a large amount of data available for these enzymes. There are well over 2500 sequences

available in Pfam (Finn *et al.* , 2010) and more than 250 experimentally determined structures are available in the PDB (Berman *et al.* , 2000).

1.4 Cystatins

1.4.1 Summary

Cystatins are protein inhibitors of cysteine proteases. They bind to the protease active site and physically block the substrate from binding. The side-chains of the inhibitor binds in the binding pockets of the protease but the backbone of the inhibitor turns away from the active site where cleavage would normally occur (Figure 1.2). They are used to prevent uncontrolled degradation of valuable proteins by endogenous cysteine proteases and also, in some cases, to inhibit the cysteine proteases of other species in inter-species interactions. Since the interaction between the inhibitor and the protease is non-covalent and the inhibitor binds to the active site, the inhibition is classified as competitive, reversible inhibition.

1.4.2 History

The first cystatin to be described was ovocystatin, isolated from chicken (*Gallus gallus*) egg white although at the time it wasn't called cystatin (Fossum & Whitaker, 1968). Later, similar inhibitors were discovered in mammalian tissue. The name, cystatin, was proposed in the eighties (Barrett, 1981). The structure of a ovocystatin was the first structure of a cystatin molecule to be published in 1988 (Bode *et al.* , 1988). Thereafter several more structures were published including the structures of human-derived cystatins.

1.4.3 Classification

The classification schema described here is that followed by the MEROPS database (Rawlings *et al.* , 2010). The process for classifying the inhibitors was described in an earlier paper (Rawlings *et al.* , 2004). Protease inhibitors are classified into clans and families based on homology. Homology is inferred based on structural or sequence similarity. Inhibitors with similar tertiary structure are grouped together in a clan. The clan is then further divided into families based on sequence homology. Clans are named with the the letter "I" or "J" ("I" for inhibitor) followed by another letter. The reason for using "J" is because there are more than 26 clans and thus to be able to name them all another letter apart from "I" was needed. "J" was chosen because it is the first letter in the alphabet after "I". Families are indicated with an "I" followed by a number. As with proteases the number is unique to that family across all clans. Some proteins contain more then one cystatin domain. In some cases these domains are not from the same cystatin sub-family. As such, the classification of cystatins applies to the inhibitory domains and not to the whole protein.

Cystatins were given the identifier I25 and, together with the latexin family, they form clan IH.

Cystatins are classified into four subfamilies. The subfamilies are based on sequence similarity but also share many functional characteristics. Each sub-family has a single "type" protein which serves as the model for the sub-family. If a cystatin domain shows sequence similarity to that protein above a certain threshold it is classified in that sub-family. The subfamilies are listed below with accompanying "type protein":

I25A Cystatin-A (*Homo sapiens*) - Type I These are mostly endocyttoplasmic cystatins but can also be found in body fluids. They typically do not possess disulphide bonds and are not glycosylated. Human cystatin-A and cystatin-B fall into this sub-family. This sub-family is also known as the stefins.

I25B Ovocystatin (*Gallus gallus*) - Type II These are mostly excreted cystatins and contain two conserved disulphide bonds and are sometimes glycosylated. Human cystatin-C and kininogen subunits-2 and -3 fall into this sub-family. In some literature this sub-family is referred to as the cystatin sub-family, usually when referring to the sub-family I25A as stefins.

I25C Metallopeptidase inhibitor (snake venom, *Bothrops jararaca*) These are not inhibitors of C1 proteases. Human kininogen subunit-1 falls in this sub-family.

I25D Unassigned There are some members of the cystatin family that could not be assigned to one of the other sub-families and they have been put here for the time being.

For simplicity, the names "Type I" and "Type II" will be used in the rest of the study for the families indicated above.

1.4.4 Structure and function

See Figure 1.2 for a visual representation of cystatin interacting with a protease. Cystatins contain a four-strand β -sheet with an α -helix lying across it, perpendicular to the direction of the strands. Cystatin binds to a C1 protease via three distinct binding regions. These regions are made up of the N-terminal and two loops between the strands of the β -sheet. The inhibitory function is achieved by binding to the protease active site and physically blocking access to it. The backbone turns away from the protease in the place where a substrate protein would normally be cleaved. As such there is no labile bond in the position where cleavage would normally take place. Cystatins are reversible, competitive inhibitors since the interaction is temporary and competes with the substrate for access to the binding site. Although the interaction is reversible, it is a tight-binding inhibitor with inhibition constants, typically, in the nanomolar range.

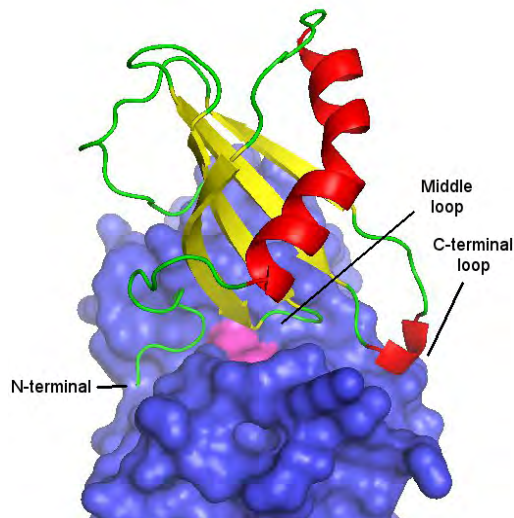


Figure 1.2 – The molecular structure of cystatin-B (PDB ID: 1STF) shown in cartoon representation. The protease (papain) is shown in surface representation in blue with the catalytic cysteine in purple. From the N-terminal, the backbone of the inhibitor fits in the active site cleft but turns away from the protease at the catalytic cysteine, rejoining the cleft on the other side of the catalytic residue.

1.4.5 Biological roles

Cystatins play a crucial role in preventing proteolysis by C1 proteases; therefore their biological role often corresponds with that of the proteases they inhibit. Type I cystatins are often used intracellularly to prevent proteolysis by proteases that accidentally leak from their sub-cellular compartments (Turk *et al.*, 2002). Type II cystatins are mostly secreted into bodily fluids and plasma. They often function in the protection against micro-organisms that use C1 proteases to invade the body. They also regulate the function of C1 protease that operate extracellularly, for example, cathepsin-B in the process of bone resorption. The role of the cystatin inhibitory units on kininogen is not clear yet. One well documented function of phytocystatins is the inhibition of C1 proteases in the gut of herbivore insects. This prevents the digestion of plant proteins by the insect; thereby protecting the plant. The cystatins, although reasonably well-studied are not understood as well as the proteases that they inhibit.

1.4.6 Clinical significance

Cystatins are important in various diseases and clinical applications. Some examples are given below:

- Plasma levels of Cystatin-C serves as an indication of glomerular filtration rate in the kidneys. This is used as a diagnostic for kidney damage (Mussap & Plebani, 2004).

- The interaction between cystatin-C and cathepsin-B plays a role in the spread of many types of cancer (Mussap & Plebani, 2004).
- Certain mutations in cystatin-C lead to amyloid formation and cerebral hemorrhage (Abrahamson, 1988).
- Mutations in cystatin-B are thought to be responsible for progressive myoclonus epilepsy (EPM1) (Pennacchio *et al.* , 1996).
- Imbalances between cystatins and cysteine proteases play a role in various diseases including osteoporosis, rheumatoid arthritis and Alzheimer's (Vasiljeva *et al.* , 2007).
- Certain parasitic nematodes use cystatins to modulate of the host immune response (Hartmann & Lucius, 2003).

Unlike with the C1 proteases there have been no attempts to target cystatins with drugs. The clinical application of cystatins is currently limited to diagnostics.

1.4.7 Economical significance

The industrial use of cystatins is currently limited. There is, however, some interest in cystatins in the biotechnology sector in agriculture.

Phytocystatins play an important role in defense against some herbivorous insects. After ingestion by the insect, the inhibitor blocks the digestive enzymes of the insect in the insect gut. This leads to reduced growth or death for the insect. Much has been published on the topic recently (Martinez & Diaz, 2008). There are various attempts to utilize this process in biotechnology to make crops more resistant to pests.

1.4.8 Available data

The cystatins are a well-known and well-documented group, having been studied for over 40 years. There are over 950 unique sequences of cystatins in the Pfam (Finn *et al.* , 2010) database and over 33 experimental structures in the PDB (Berman *et al.* , 2000).

1.5 Chagasin-like inhibitors

1.5.1 Summary

Chagasin-like inhibitors are a recently discovered family of protein inhibitors of C1 proteases that was first identified in *Trypanosoma cruzi* (Monteiro *et al.* , 2001). Homologs were later

identified in many different organisms ranging from *archaea*, *eubacteria* and *eukaryotes* (Rigden & Mosolov, 2002). The only known molecular function is the inhibition of C1 proteases. Very little is known about the actual physiological processes they are involved in. Since the protein family is widely distributed amongst different organisms, implying a long evolutionary time-period, it is likely that the members have adapted to play a variety of physiological roles. Several experimental structures have been published and chagasin-like inhibitors share a striking similarity to cystatins in the regions interacting with the protease.

1.5.2 History

The discovery of chagasin was announced in 2001. In that same year, based on secondary structure prediction and threading, it was correctly postulated that chagasin has an immunoglobulin-like fold (Monteiro *et al.* , 2001). The discovery sparked strong interest in the research community. Soon after, in 2002, a list of homologs discovered in other species were published; along with an analysis of the conserved residues and possible mechanism of interaction with the protease (Rigden & Mosolov, 2002). The first experimental structure was an NMR structure published in 2006 (Salmon *et al.* , 2006). A monomeric crystal structure (Da Silva *et al.* , 2007) and several crystal structures of chagasin in complex with different proteases followed (Ljunggren *et al.* , 2007; Wang *et al.* , 2007; Redzynia *et al.* , 2008).

1.5.3 Classification

As a protein inhibitor of a protease, chagasin-like inhibitors are classified in the same schema as the cystatins. See subsection 1.4.3 for an explanation of the classification schema. Chagasin-like inhibitors have been assigned to clan JL and family I42. There are no subfamilies defined.

1.5.4 Structure and function

Chagasin has an immunoglobulin-like fold. It has two β -sheets, each containing three β -strands, lying on top of each other. As with cystatin, the interacting parts lie in the loops that connect the secondary structure elements. The inhibition mechanism is strikingly similar to that of the cystatins. Three separate loops bind to the active site groove in exactly the same places as the three loops of the cystatins. The inhibitor physically blocks access to the active site. Like the cystatins, chagasin-like inhibitors are tight-binding, reversible, competitive inhibitors. See to Figure 1.3 for a visual representation.

1.5.5 Biological roles

Not much is known apart from the fact that chagasin-like inhibitors are inhibitors of C1 proteases. There are likely a wide variety of functions, based on the wide distribution of the protein family and the wide variety of functions of C1 proteases.

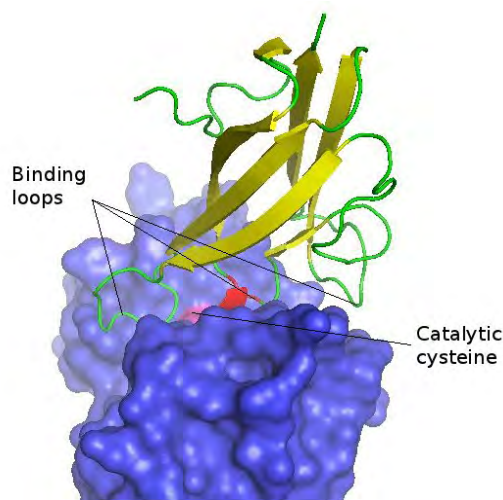


Figure 1.3 – Molecular structure of chagasin (PDB ID: 3E1Z) in cartoon representation. The protease (papain) is shown in surface representation in blue with the catalytic cysteine in purple. Chagasin is shown in cartoon representation. As with cystatin, the backbone of the inhibitor turns away from the protease at the catalytic cysteine residue.

1.5.6 Clinical significance

Chagasin-like inhibitors are present in many bacterial and protozoan pathogens. This includes *Trypanosoma*, *Leishmania*, *Entamoeba*, *Pseudomonas*, *Clostridium*, *Coxiella* and *Plasmodium spp.* (Rigden & Mosolov, 2002; Pandey *et al.* , 2006). Some specific examples of involvement in disease are chagasin itself, shown to be a virulence factor in Chagas disease (Santos *et al.* , 2006), and falstatin, a distant homologue of chagasin from *Plasmodium falciparum*, shown to facilitate erythrocyte invasion during malaria (Pandey *et al.* , 2006). Because of the important roles of many C1 proteases in pathogenic diseases it is thought chagasin-like inhibitors may also have important functions. A chagasin homolog has not been identified in humans. Currently there are no clinical uses for chagasin-like inhibitors.

1.5.7 Economical significance

It was shown that, similarly to cystatin, chagasin can inhibit the digestive enzyme of bean weevil larvae and cause mortality and reduced growth amongst them (Dos Santos Monteiro *et al.* , 2008). This is aimed at the creation of genetically modified, weevil resistant, crops. Further than that, the known economical significance of these proteins are limited to the cost of the diseases they are involved in.

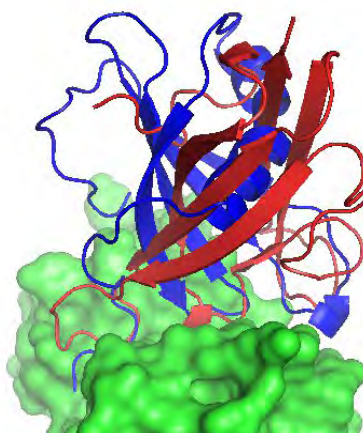


Figure 1.4 – Comparison of the molecular structures of cystatins and chagasin-like inhibitors. Cystatin-B (PDB ID: 1STF) is shown in blue and chagasin (PDB ID: 3E1Z) in red, both in cartoon representation. The protease, shown in green, is papain from 1STF. The orientation of chagasin to the 1STF structure is based on the superpositioning of the proteases in the two experimental structures (papain in both).

1.5.8 Chagasin-like inhibitors of malarial proteases

In 2006 a paper was published, identifying a very remote homolog of chagasin in *P. falciparum* (Pandey *et al.* , 2006). The name, falstatin, was given to the protein and it was shown to be required for erythrocyte invasion, a critical step in the parasite lifecycle. The exact function and interaction partner of falstatin remains unknown but it evidence suggests that it interacts with cysteine proteases of the host parasite to prevent apoptosis (Rennenberg *et al.* , 2010). Falstatin is much bigger than chagasin and contains N-terminal signal sequences and additional inserts.

1.5.9 Available data

The chagasin-like inhibitors have only been discovered recently (2001), yet there are already roughly 170 sequences in Pfam (Finn *et al.* , 2010) and 10 structures in the PDB (Berman *et al.* , 2000). While not as much data as for the cysteine proteases or cystatins, this is enough for the purposes of this study.

1.6 Comparison between cystatin and chagasin

Cystatin and chagasin are both protein inhibitors of C1 proteases. The overall structure of the cystatins and the chagasin-like inhibitors differ completely but there is almost perfect structural similarity between the interacting loops of the two proteins. See Figure 1.4 for a

visual comparison of the two inhibitor families. The similarity between the binding loops of cystatins and chagasin-like inhibitors is a truly remarkable case of convergent evolution. The function of the proteins are carried out solely by these interacting loop and the rest of the protein serve only to hold these loops in the correct orientation. Thus, two different scaffolds evolved independently, yet perform the same function using the same method.

Cystatins are more widely distributed amongst different phyla than chagasin-like inhibitors. Cystatins predate animal and plant divergence since they are found in both kingdoms. Chagasin is more narrowly distributed and has so far only been discovered in parasites of animals. Combined with the fact that chagasin has an immunoglobulin-like fold this has led to the hypothesis that chagasin arose from a horizontal gene-transfer event, from host to parasite (Rigden *et al.* , 2001).

1.7 Computer-based protein structure prediction

A protein's function depends on it's structure. Therefore, knowledge of a protein's structure is useful in understanding how a protein carries out it's function on a fundamental level. Furthermore, structural information can be used to guide biochemical experiments like mutagenic studies and reduce the amount of laboratory work that is necessary to characterize a protein (Zhang & Kim, 2003). It is also useful when designing molecules that interact with the protein. Structure based drug design is one way of obtaining drugs that target specific proteins and requires structural knowledge of the drug-targets. Although most drugs are still derived from natural sources, it can be argued that as the drug-design techniques improve, it will become more economical to use structure-based drug design. There is hope that the usefulness of knowing a protein's structure will improve over time as the tools to make use of the information improves (Andricopulo *et al.* , 2009).

Determining the 3D structure of a protein through experimental procedures is expensive, although the structural genomics project has reduced the cost somewhat (Pusey *et al.* , 2005). There are currently two main technique of experimental structure determination that account for over 99% of experimentally determined protein structures.

The first and oldest technique uses X-ray diffraction (Kendrew *et al.* , 1958) to infer the structure of a protein. The protein is crystallized and then used to diffract X-rays from an X-ray source. The diffraction pattern is then used to infer the structure of the protein via complex mathematical equations.

The second technique uses nuclear magnetic resonance (NMR) to create a set of spatial constraints used to solve the structure of the protein (Wuthrich, 1989). NMR refers broadly to the use of magnetic fields and electromagnetic radiation to detect energy shifts caused by interactions between atomic nuclei and electrons in the protein due to their magnetic spin. The results are obtained as a spectrum of frequencies where the protein interacts with the electromagnetic radiation. These clues are then used to reconstruct the protein structure.

A third technique, utilizing electron microscopy, is gaining traction and is mainly used for getting low resolution structural information about very large proteins. As there are no cysteine protease structures or any of their inhibitors determined with this technique, it will not be discussed further.

Both X-ray diffraction and NMR have important limitations. In the case of X-ray diffraction, purification of sufficient protein and subsequent crystallization is difficult and sometimes impossible with very large proteins. Another problem is that a structure obtained through the use of X-ray diffraction represents the protein conformation in a crystallized form; which is often not the same as that of the functional protein in physiological conditions. With NMR, the main problem is that interpretation of the raw NMR data is very difficult. In large molecules it becomes impossible to reconstruct the protein structure from the spectrum. As such, NMR is limited to small proteins.

Computer-based structure prediction refers to the theoretical prediction of protein structures based on their sequence, using the computational power of computers. Since it was shown that a protein's sequence determines its 3D structure (Anfinsen & Haber, 1961), attempts have been made to unravel this causal relationship. Computer-based structure prediction is the only viable method addressing the problem.

There are two main branches of the discipline namely *de novo* (also called *ab initio*) modelling and homology modelling. Software can be classified into one of these branches based on the information that it uses to make the structural prediction. If the program uses experimentally derived structural information about a homologous protein, then it is classified as homology modelling. Otherwise it is considered *de novo*. Apart from homologous protein structures, several other sources of information are incorporated into the predictions. This includes physics based energy functions and force fields as well as statistical potentials derived from the knowledge base of all solved protein structures (Moult, 2005).

All computer-based structure prediction contains two important elements, namely conformational search and conformational evaluation. No matter what the algorithm, these two elements are always present. Viewing computer based structure prediction programs within this paradigm is very useful as it makes comparisons between different programs and techniques sensible.

The average protein contains thousands of atoms and each atom has three spatial dimensions. Thus, there are typically thousands of variables involved when specifying a protein structure. Proteins thus have an extremely large range of theoretically possible conformations (conformation space). Effectively sampling this huge space and reliably evaluating the sampled conformations in terms of their thermodynamic favorability is, in essence, what all computer based structure prediction software tries to do.

To monitor and evaluate the software used to produce structural models of proteins, a major project was created in the form of a structure prediction competition called "Critical Assessment of Techniques for Protein Structure Prediction" (CASP) (Moult *et al.*, 1995). Proteins whose structures have been experimentally determined, but not yet published, are designated as targets for computer based prediction. Entering teams then submit predictions for some

or all of the target proteins. The submitted models are then compared to the experimental structures to determine which models were the best. The competition has been running since 1994 and is held every second year. A website dedicated solely to this competition can be found at www.predictioncenter.org. CASP is the main source of information comparing the accuracy of different modelling techniques and programs.

The results of CASP show that *de novo* modelling is steadily improving but cannot currently provide reliable results for anything but very small proteins (Ben-David *et al.* , 2009). There are still insufficiencies in both the conformational search, and the identification of native conformations amongst the sampled conformations. Although the use of special-purpose hardware like Anton (Shaw *et al.* , 2007) can dramatically speed up molecular dynamic simulations, access to such computers is limited and expensive. Homology modelling, on the other hand, can produce accurate, reliable and cheap results on commodity hardware, but is limited by the availability of experimentally determined structures of homologous proteins (Kryshtafovych & Fidelis, 2009).

1.8 Homology modelling

1.8.1 Introduction

Homology modelling refers broadly to the practice of creating a structural model of a protein (target) by using experimentally obtained, structural, information about a homologous protein (template). The idea dates back to the early 80's, before the wide-spread use of computers, when it was pioneered with mammalian serine proteases using physical models (Greer, 1981). It relies on the conservation of protein structure during evolution (Chothia & Lesk, 1986). Finding sequence similarity between proteins practically guarantees that there will be structural similarities. The structural information allows the program to limit the conformational search to conformations that resemble the homologous protein, reducing the search space immensely. Depending on the evolutionary distance between the target and template, highly accurate models can be produced (Read & Chavali, 2007). Some have even been used for drug-design purposes (Hillisch *et al.* , 2004).

Like with the experimental methods, there are certain limitations to using homology modelling. Homology modelling can only be used if the structure of at least one closely related protein is known. Furthermore, it is difficult to know how accurate a model is after it is built, although a good estimate can be made based on the sequence similarity to the template.

Homology modelling increases information recovered from the experimental methods. By determining the structure of just one protein it is actually possible to get information about the structures of many, related, proteins. It is hoped that homology modelling will be used, in the near future, to provide reliable structural information about all sequenced proteins and close the gap between sequence and structure data (Zhang & Skolnick, 2005).

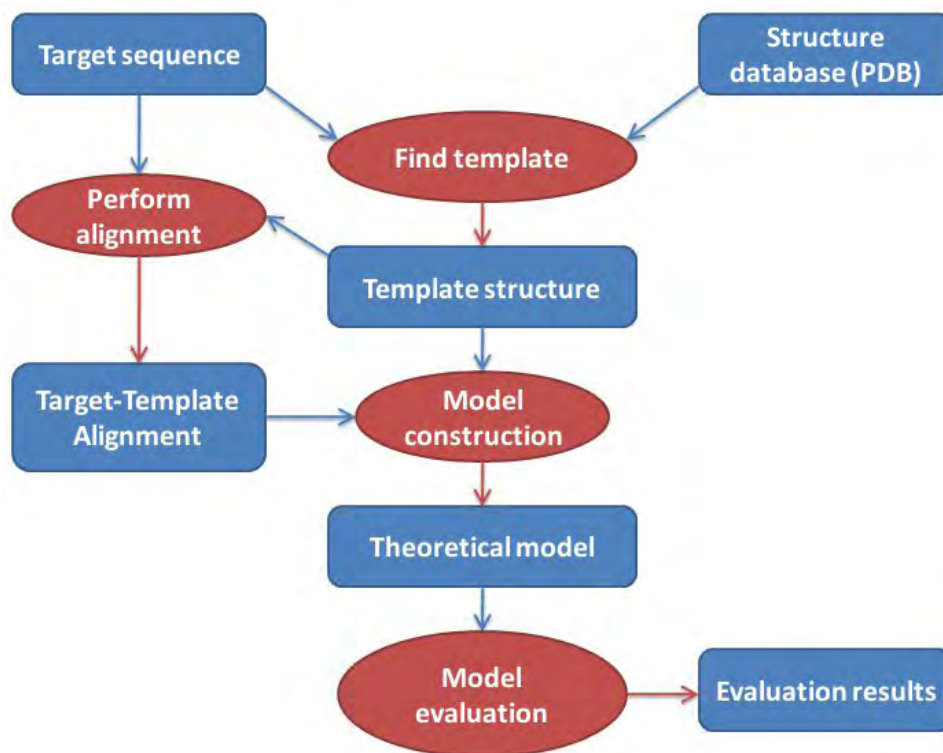


Figure 1.5 – Flow diagram of homology modelling process. Data is shown in blue rectangles and processes are shown in red ovals. Blue arrows represent input into a process and red arrows represent output.

1.8.2 Process overview

The process described here is the traditional homology modelling approach. Many variations exist, but the process outlined here is still a useful representation of the process.

The protein being modelled is called the target. The related protein, or proteins, with experimentally determined structures are called the template, or templates, because its structure serves as a template for that of the target protein. The method involves four main steps (Moult *et al.* , 2007; Sanchez & Sali, 1997; Xiang, 2006):

(See Figure 1.5 for a graphical presentation)

1. Homolog identification: The first step in applying homology modelling is to find suitable experimental structures on which to base the new model. This step relies on the existence of a homolog with known structure as well as the ability to find it. Various search algorithms have been implemented in computer programs and are used for this purpose. These are the same as those used to detect homology in other research applications. Currently the most widely used search tools are based on the BLAST algorithm and variations of it (Altschul *et al.* , 1997). Depending on the specific method

that will be used to construct the model, one or more homologs are chosen to serve as templates. Structures are mostly extracted from the PDB database (Berman *et al.* , 2000).

2. Alignment: The sequence of the target protein is aligned to that of the template protein or proteins. Alignments can be performed by submitting sequences to a program or webserver that return sequence alignments (Pei *et al.* , 2008; Notredame *et al.* , 2000). There have been many recent improvements in alignment software and manual editing of alignments is usually not required. The alignment step is critical to the success of a homology modelling attempt. Any errors in the alignment of target and template are likely to lead to large errors in the final model. Alignment of distant homologs can be improved by using multiple sequence alignments, Hidden Markov Models (HMM), profile alignments or using structural information (Jimin, 2008).
3. Model creation: The model is built by submitting the alignment of the target and template protein, from step 2, to a modelling program along with the structure of the template protein or proteins. The program then constructs a model of the target protein. Different programs have different methods of incorporating the information from the homologous protein structure. There are also different methods of searching the conformational space and evaluating individual conformations. Sometimes they have several evaluation functions that are used in different stages of the model building process. Physical energy, stereo-chemistry, similarity to known protein structures and similarity to the template are the main criteria used in the evaluation functions. Putative model generation techniques include steepest descent optimizations, genetic algorithms and Monte Carlo techniques (Sali & Blundell, 1993; Petrey *et al.* , 2003; Wallner & Elofsson, 2005; Misura *et al.* , 2006; Xiang, 2006).
4. Model validation: The model is evaluated using several tools to try and find any errors and estimate the accuracy of the model. Different tools may be used to analyze different aspects of the model. The evaluation tools usually give a score that indicates how "normal" a model is compared to known protein structures. For example, bond-lengths, bond angles, angle combinations and inter-atomic packing are all parameters that may be evaluated (Kryshtafovych & Fidelis, 2009). If errors are found, they may be corrected and re-evaluated. Step 4 is important and can also be used to select a model from a set of putative models. Different applications of protein models require different levels of accuracy. It is therefore important to assess the accuracy of models so that they are not used for a purpose for which they are not accurate enough. The accuracy of homology models can vary from completely wrong (ie, wrong fold) to highly accurate models, almost as good as an experimental structure. The true level of similarity between the model and true protein is best estimated from the sequence similarity between the template and target proteins (Kryshtafovych *et al.* , 2011). Information about the divergence of sequence and structure (Chothia & Lesk, 1986) is very useful in making this estimation. There is ongoing efforts to try and find better ways of estimating model accuracy (Benkert *et al.* , 2011).

1.8.3 Modeller

Modeller was the program chosen for performing homology modelling in this project. The reasons for choosing modeller above other packages are mainly speed, a proven track-record and ease of use.

Modeller (Sali & Blundell, 1993) is a program that is very commonly used to produce homology models. In over seventeen years of existence, it has been continuously updated and expanded. Today, it is a structural biology toolbox that also provides functions for sequence alignment, structural superpositioning and more. Regular updates are released and there are very few bugs in the program. It is free to use to academic and non-profit institutions but requires registration (free of charge) to download the program. Modeller can be obtained from <http://salilab.org/modeller>.

Modeller uses optimal satisfaction of spatial constraints to produce homology models. Spatial constraints are represented as probability density functions (pdf) to describe a specific spatial feature. A pdf indicates the probability that a specific variable will have a certain value. Different pdfs will put different constraints on the variable. By applying various pdfs to all the atoms in the protein, certain patterns and rules can be enforced on the position that the atom is allowed to take.

By examining known protein structures, various such pdfs were determined for many features within proteins. To apply certain pdfs to only specific atoms, conditional pdfs were implemented. Conditional pdfs are only applied if the conditions attached to it are met. For example a conditional pdf will provide information about a bond-length only if the two atoms involved are both carbon. Each conditional pdf may have multiple conditions, combined with Boolean logic. For example instead of treating all carbon-carbon bonds the same, a pdf can be used that applies to a specific carbon-carbon bond. For instance the bond between the α -carbon and the carboxyl carbon will have its own pdf and the bond between the α -carbon and the sidechain β -carbon will have another. There are many characteristics that may be described by pdfs, some examples that are used by Modeller are bond-lengths, bond-angles, dihedral angles, and non-bonded inter-atomic distances.

To use information from the template structure, some characteristics of its structure are converted into pdfs. Information from many known protein structures are combined with that from the template to do this. This is because it is impossible to generate a pdf from a single observation. Information about distribution shapes and variances from large numbers of proteins is combined with the single observation from the template to create each pdf.

Features in the protein are described by more than one basic pdf and these are combined by weighted summation or multiplication, where appropriate, to give feature pdfs. The feature pdfs listed in the article are:

- C α -C α distance restraints
- Main-chain N-O distance restraints

- Stereo-chemical restraints
- Main chain conformation restraints
- Side-chain dihedral angle restraints

These exist for every instance of the feature in the protein separately.

To find a model that optimally satisfies all the pdfs, the feature pdfs are combined into one single molecular pdf. This function is a multidimensional ($3N$ dimensions where N is the number of atoms) scalar function. The values or coordinates that maximizes this function corresponds to the coordinates of the atoms in the model that optimally satisfies all the feature pdfs. To combine the feature pdfs into one molecular pdf, they are simple multiplied with each other. This gives the function that is to be optimized.

An iterative approach (where the conformation of the protein is repeatedly adjusted) is used to optimize the function where at first only very short range terms are optimized and the successive steps increases the range of terms included until all the terms are included. The optimization uses another transformation: $F = -\log(P)$. The function F is called the objective function and has a minimum where the molecular pdf has a maximum. However the function F is more friendly to computations since the log function transforms product terms into sum terms.

A conjugate gradient method is used to find the minima during each step, using the answer from the previous step as the starting point for the next step. After all the steps are complete and the molecular pdf have been optimized, the point in multidimensional space that optimizes the molecular pdf defines the structure of the protein. From this point, further refinements to the protein structure are made, depending on the options specified by the user. This can include short molecular dynamics runs and simulated annealing.

1.9 Protein-protein interactions

Proteins often act together by binding to each other. These interactions ranges from transient interactions to permanent binding for the lifetime of the protein. The binding of one protein to another can change the functional activity of the proteins. This change can be as small as a slight change in the kinetics of an enzyme or as big as changing the catalytic reaction of the enzyme completely. Some enzymatic functions can only be carried out when several proteins combine to form a multi-protein complex. The same forces that govern protein folding also govern those of protein-protein interactions. Hydrophobic interactions are indicative of permanent associations between proteins whereas hydrogen bonds and electrostatic interactions indicate temporary association often dependent on environmental conditions such as pH (Jones & Thornton, 1996). There are various ways to predict the structures of protein-protein complexes. The term docking is often used for the process of predicting structures of protein complexes. For the sake of clarity, in this document, the phrase homology modelling will

be used to refer to the prediction of complex structures using information from the known structure of an evolutionary related complex. The term, docking, will be used for all other methods of predicting protein complex structures (this includes methods that make use of homologous monomer structures).

1.9.1 Docking

Docking is a term used broadly for the prediction of binding conformations of ligands to proteins. In the case where the ligand is a protein it is often referred to as protein-protein docking. There exists a vast array of different techniques and approaches for docking (Vakser & Kundrotas, 2010). They utilize physical forcefields, statistical potentials and complementary spatial conformations, amongst other, to predict binding positions. Docking traditionally relies only on structural knowledge of the monomers that combine to form the complex. The monomeric structures are then combined through the docking process to give a complex structure. For docking to be feasible, the structures of the binding partners needs to be known quite accurately. There are many cases where homology models are sufficiently accurate to be used for docking (Tovchigrechko *et al.* , 2002). In cases where the proteins have the same conformation in the monomeric (unbound) and complex (bound) form, accurate predictions of the complex structure are generally possible. To improve the accuracy of docking attempts, and reduce the computational expense, it is helpful to use additional information to limit the number of possible conformations considered. Experimental data that implicate or exclude certain residues in the binding of the proteins is a common type of such information.

A recent paper (Vajda & Kozakov, 2009) identifies three main classes of docking programs based on the approach used. These are fast Fourier transforms (FFT), medium range Monte Carlo methods and restraint-guided high-ambiguity driven biomolecular docking (HADDOCK). Leading programs implementing each is ZDOCK (FFT), RosettaDock (Monte Carlo) and HADDOCK, which is also the name of a program. These methods utilize different source of information and are suited to different docking problems. FFT programs are best suited to, and perform best at, global searches, where there is no prior information about binding location is known. It uses a simplified representation of the protein surface to find complementarities in shapes of the two proteins. Monte Carlo techniques are very expensive computationally and are therefore unsuitable for global searches. They excel at optimizing local conformations and making refinements to conformations that are already close to the native structure. HADDOCK type docking excels when there is prior knowledge of the binding interface that can be used to guide the docking attempt. HADDOCK can produce good results even with significant conformational changes upon binding.

A similar project to CASP was launched for protein-protein complexes in 2001 (Janin *et al.* , 2003). The Critical Assessment of PRedicted Interactions (CAPRI) follows the same principle as CASP in setting targets with known, but unpublished, structures. Steady progress has been observed over the duration of the competition (Lensink & Wodak, 2010).

1.9.2 Homology modelling of protein complexes

Applying the homology modelling methodology to protein-protein complexes is a relatively new approach (Kundrotas & Alexov, 2006; Kundrotas & Vakser, 2010). It has exactly the same steps as the normal homology modelling method for monomers. This is a relatively new approach for complexes because the number of experimentally solved complexes has only recently become high enough to make this approach feasible in a significant number of cases (Hwang *et al.*, 2008). There is still relatively little information available in the literature regarding the efficacy of this approach. Results published so far, including from this study (Tastan Bishop & Kroon, 2011), are promising. Theoretically, homology modelling should be able to provide better results than docking, since more information is utilized to make the prediction. In fact, homology modelling can be seen as a special case of docking where information from a homologous complex is used to limit the conformational search. However, docking techniques and homology modelling techniques have followed separate development paths and the implementations of the two approaches differ markedly. Since the divide between homology modelling and docking is only artificial, from a theoretical viewpoint, it is possible that a convergence of the techniques will take place (Vajda & Kozakov, 2009). The docking methods typically spend more time in optimizing the orientation of the monomer structures relative to each other while homology modelling usually leaves the position of the backbone atoms as close to their position in the template as possible. Homology modelling is much less time consuming than docking implementations, as a much smaller conformational space is searched.

1.10 Data and databases

For the sake of convenience, most biological sequence and structure data are saved in online databases, so that the data can be obtained easily. A list of the databases used in the current project is given here.

1.10.1 UniProt (www.uniprot.org)

UniProt (One protein) is a curated database that aims to store rich information about all known proteins. Its most valuable feature is that it is non-redundant and only has one database entry for each individual protein. That makes it very useful to get reliable and up to date data on a specific protein. It provides sequence data as well as annotation for each protein including post-translational modifications, literature references and links to other databases, including the PDB (Apweiler *et al.*, 2010).

1.10.2 Protein Data Bank (PDB, www.pdb.org)

The PDB is the main database for experimentally determined protein structures. Each structure is given a four letter code (made up of letters and numbers) which uniquely identifies the structure. This code is called the PDB ID and is used throughout this study to refer to specific experimental structures (Berman *et al.* , 2000). By going to the website address given above and entering the PDB ID in the search box, the original structure and information relating to it can be found.

1.10.3 MEROPS (merops.sanger.ac.uk)

MEROPS is a database exclusively dedicated to proteases and their inhibitors. It provides a classification schema and assigns clan and family names. It also has a list of members of each family with links to other databases and literature (Rawlings *et al.* , 2010).

1.10.4 Pfam (pfam.sanger.ac.uk)

Pfam (Protein family) provides alignments of protein families. It also provides phylogenetic trees showing the distribution of the members of each family in the living kingdoms. Pfam draws its sequences from the UniProt database. New sequences are regularly and automatically added to the existing families. Because it is automated, it is very up to date (Finn *et al.* , 2010).

1.11 Summary

The C1 protease family and its protein inhibitors are relevant research topics. These proteins are important for medical research, agriculture and even biochemistry research itself. The C1 proteases are also a very well characterized protein family with many sequences and experimental protein structures available. This large amount of experimental data enables computational methods to be used that are not feasible for most protein families. One specific computational approach, that is explored further in this project, is the use of homology modelling to predict the 3D structures of protein-protein complexes. This has the advantage, over conventional docking procedures, of much lower computational cost while maintaining very high accuracy.

Chapter 2

Project design and rationale

2.1 Chapter overview

This chapter aims to give the reader insight into the overall rationale and experimental design of the project. It also provides technical details about the protocols and programs that do not fit into one particular part of the project but was important throughout. Some terminology is also introduced to simplify discussion later on. The project aims and motivation are given first (Section 2.2), as this provides context to the rest of the chapter. The experimental design of the entire project is then described without much technical details (Section 2.3). This is followed by technical details of the homology modelling done in the project, including the calculation of scoring functions and energy minimization (Section 2.4). The chapter ends by providing discussions of technical considerations of the data management, software, used in the background, throughout the project and computational resources used. (Section 2.5 and 2.6).

2.2 Project aim and motivation

The main research question addressed in this study was whether templates constructed by combining coordinates from different experimental structures, using superpositioning, can be used to construct accurate models of C1 protease-inhibitor complexes.

This question was addressed by generating such templates and using them in a homology modelling pipeline and comparing the results to those achieved by using traditional templates in the same pipeline. The accuracy of both of sets of models were estimated by using established accuracy metrics. Various techniques to improve model quality were used both to evaluate their effectiveness and to evaluate their suitability for models based on the two types of templates.

Additional research questions addressed are whether C1 proteases and/or their protein inhibitors undergo conformational change upon complex formation, how to select the best templates for homology modelling of C1 protease-inhibitor complexes and how accurate the models are.

The original aim, when the project started, was to accurately model, and then analyze, several complexes of cysteine proteases and protein inhibitors that were of interest due to their role in disease and biotechnology. It became clear that if this was to be done for a large number of complexes, homology modelling would be the best technique to use. Although not well established for the use of predicting protein complex structures, homology modelling is a better choice than traditional, monomer-based, docking techniques in this case. The reason is that there are several experimentally determined complex structures of cysteine proteases with protein inhibitors available. Given these structures, the matter of orientation between monomers become a trivial issue, and the main focus of the modelling attempt must be the correct prediction of the monomer structures and the refinement of the interface regions. While medium range Monte Carlo docking methods and HADDOCK (Section 1.9.1) could also be used for this purpose, homology modelling has the added benefit of very low computational requirements and easy incorporation into automated protocols.

Once the decision to use homology modelling was made, the next question was which templates to use. There are well studied criteria for the modelling of monomers, but for complexes there was little information to be found regarding how to prioritize the sequence similarity of the separate monomers when a choice of template was to be made. The problem could be avoided altogether if the best template for each component of the complex could be selected from different experimental structures. This removes the need to compromise on the sequence similarity of the individual proteins. A very simple way to do this is to superimpose the experimental structures onto each other, based on the common parts, and then copy the coordinates for the individual components from the superimposed version of the structures. In this way, a hybrid template is formed that contains components from two separate experimental structures. This has not been done before and, as such, the approach carries some risk as the reliability of such templates has not been verified in previous studies.

In order to answer the question of which templates were the most suitable and whether hybrid templates could be used, a separate study was launched that became the main focus of the project. The aim of this study was to directly compare the accuracy of models based on hybrid templates to those based on single-experimental-structure templates (simple templates) and elucidate the relationship between the sequence similarity of the monomers and the accuracy of the complex model. As part of this investigation, the effect of energy minimization to improve the accuracy of models and the ability of scoring functions to identify accurate models were also investigated. The modelling of a large number of cysteine protease complexes with protein inhibitors was still carried out, but, due to time constraints, the analysis of these models was scaled down from the originally intended analysis.

2.3 Experimental design

The project comprises four main parts. Three parts are experimental in nature, with the fourth forming part of the results presentation. The first part involves the examination of experimentally solved protein structures. The second involves the modelling of complexes with known experimental structures to evaluate the accuracy of models built on different templates. The third follows from the second, and uses its results to select templates for, and model, several novel complexes. The last part consists of the creation of a website to provide access to all the raw data of the project in a useful way.

2.3.1 Part 1: Examination of experimental structures

In the first part, some experimental structures (Table 3.1) of all the protein families involved in the project were investigated. One of the major aims of this study was to test if any of the proteins undergo conformational changes upon binding. Such conformational changes would have very important implication for the selection of templates, as conformational changes are not accurately modelled using homology modelling. The other major aim was to examine the variation that exists between different crystal structures as a measure of what accuracy could realistically be achieved using homology modelling. Calculation of RMSD values and visual inspection was used as the main investigative techniques. The study also provides some information about the differences between the subfamilies that was beyond the scope of the introduction. A detailed description of the methodology followed is given in Chapter 3.

2.3.2 Part 2: Construction and evaluation of test set models

In the second part of the project, many models were built, and evaluated, in order to address the issue of template selection for the modelling of complexes as explained in Section 2.2. To directly compare the accuracy of different models, an objective measure of accuracy is needed. To accomplish such objective measurement, this part of the project was modelled on the CASP and CAPRI competitions. Models of complexes, with known experimental structures, were built and then compared to the experimental structures using objective metrics used in CASP and CAPRI. The effect of energy minimization using Rosetta (www.rosettacommons.org) all atom minimization and the ability of two different scoring functions (Rosetta energy and DOPE Z-score) to identify high-accuracy models were also investigated. The reason for the evaluation of the scoring functions is that they offer important benefits to the homology modelling process. In most homology modelling protocols, several varying models are produced because of differing random numbers used in their construction. In a high-throughput setting, scoring functions are a viable selection tool that can be used to select the final model out of a set of candidates, thereby giving a more accurate final model.

Table 2.1 – Experimental structures used as the test set targets. These structures were also used for the generation of templates. Shown is the PDB ID as well as the name and species of origin for the protease and inhibitor of each structure.

PDB ID	Protease	Organism	Inhibitor	Organism
1NB5	Cathepsin-H	<i>S. scrofa</i>	Cystatin-A	<i>H. sapiens</i>
1STF	Papain	<i>C. papaya</i>	Cystatin-B	<i>H. sapiens</i>
1YVB	Falcipain-2	<i>P. falciparum</i>	Ovocystatin	<i>G. gallus</i>
2NQD	Cathepsin-L	<i>H. sapiens</i>	Chagasin	<i>T. cruzi</i>
2OUL	Falcipain-2	<i>P. falciparum</i>	Chagasin	<i>T. cruzi</i>
3CBJ	Cathepsin-B	<i>H. sapiens</i>	Chagasin	<i>T. cruzi</i>
3E1Z	Papain	<i>C. papaya</i>	Chagasin	<i>T. cruzi</i>
3IMA	Papain	<i>C. papaya</i>	Tarocystatin	<i>C. esculenta</i>
3K9M	Cathepsin-B	<i>H. sapiens</i>	Cystatin-A	<i>H. sapiens</i>

The effect of minimization was also studied since it offers the possibility of further improving on model accuracy.

The PDB was searched for experimental structures of complexes between C1 proteases and cystatins or chagasins that could be used as targets for the test set models. Nine such complexes were identified using keyword and BLAST searches. These structures can be viewed in Table 2.1. Each of these targets were then modelled and evaluated, using all available templates as described in section 2.4, to determine which templates gave the most accurate models. The effect of minimization on the accuracy of the models were then investigated by applying Rosetta full atom minimization to all the models and re-evaluating them using the same procedure as before. For details of the evaluation and analysis of the results, see Chapter 4. The models built in this part of the study, shall be referred to as the test set. Members of this set are referred to as test models.

2.3.3 Part 3: Construction and evaluation of study set models

In the third part of the project, ten C1 proteases and eleven protein inhibitors were selected, mostly from *P. falciparum*, *P. vivax* and *H. sapiens* (Table 2.2). These proteins were modelled, in an all against all fashion, using template selection guided by the results of the evaluations set analysis. Energy minimization and the calculation of scoring functions were also carried out on all of the models as described in section 2.4. The set of all models produced in this part of the project is referred to as the study set. Members of the set are called study models.

Table 2.2 – Proteins chosen for modelling in the study set (study set proteins). *A large portion of the falstatin sequence, including inserts near the binding loops had to be removed since there are no templates available for these regions.

Study set			
Uniprot Accession	Name	Organism	Abbreviation
Proteases			
P07858	Cathepsin-B	<i>H. sapiens</i>	Cath-B
P09668	Cathepsin-H	<i>H. sapiens</i>	Cath-H
P07711	Cathepsin-L	<i>H. sapiens</i>	Cath-L
P25774	Cathepsin-S	<i>H. sapiens</i>	Cath-S
Q9N6S8	Falcipain-2	<i>P. falciparum</i>	FP-2
Q8I6U5	Falcipain-2p	<i>P. falciparum</i>	FP-2'
Q9NAW4	Falcipain-3	<i>P. falciparum</i>	FP-3
Q6J131	Vivapain-2	<i>P. vivax</i>	VP-2
Q7Z0B2	Vivapain-3	<i>P. vivax</i>	VP-3
Q5IZD8	Vivapain-4	<i>P. vivax</i>	VP-4
P00784	Papain	<i>C. papaya</i>	Papain
Inhibitors			
Q966X9	Chagasin	<i>T. cruzi</i>	Chag
Q2PZB1	Falstatin*	<i>P. falciparum</i>	Falst
P01040	Cystatin-A	<i>H. sapiens</i>	Cyst-A
P04080	Cystatin-B	<i>H. sapiens</i>	Cyst-B
P01040	Cystatin-C	<i>H. sapiens</i>	Cyst-C
P01038	Ovocystatin	<i>G. gallus</i>	Ovocyst
Q8L5J8	Tarocystatin	<i>C. esculenta</i>	Tarocyst
P01042	Kininogen-1	<i>H. sapiens</i>	Kini-1
P01042	Kininogen-2	<i>H. sapiens</i>	Kini-2
P01042	Kininogen-3	<i>H. sapiens</i>	Kini-3

2.3.4 Part 4: Publication of results in website

In the fourth, and last, part of the project, a website was constructed in order to provide access to all the raw data produced in the study. Since a large number of models were produced and lots of metrics calculated, the resultant amount of data is large and allows for wider analysis than allowed by the scope of this project. To enable other researchers to use the data, it was made available online at <http://rubi.ru.ac.za/cpmdb>. For more details about the website and its development, please see chapter 6.

2.4 Modelling overview

This section provides an overview and technical details of all homology modelling carried out in the project. Figure 2.1 provides a graphical overview of the whole modelling process described in this section.

2.4.1 Data retrieval

The data needed for homology modelling is the sequences of the target proteins and the sequences and structures of the template proteins. In the case of the test set, the structures downloaded from the PDB, listed in Table 2.1, provides the target sequence information. The sequence information was extracted from the structure files using a custom program (Section 2.5). The sequences of the study set proteins (Table 2.2) were downloaded from UniProt. For both sets of models, the nine complex structures identified in Table 2.1 served as the source of template structure information and are referred to as the experimental template structures. The experimental template structures were identified in the PDB by manual searches using keyword and BLAST queries. All complexes found that included a C1 protease and a protein inhibitor were included and were downloaded directly from the PDB. The sequences of the study set proteins, shown in Table 2.2, were obtained by using keyword searches of the UniProt database. The sequences were downloaded directly from UniProt.

2.4.2 Sequence alignment

Multiple sequence alignments of all the protein families involved were constructed using the Promals 3D webserver (Pei *et al.*, 2008). One alignment was produced for the proteases, one for the cystatins and one for the chagasin like inhibitors. The sequences of all the proteins in the test set as well as the study set were included in this alignment. The alignment was aided by also providing protein structures from each sub-family to the webserver. As the alignment was expected to be highly accurate, given the availability of many intermediate sequences as well as structural information, it was decided that the same alignment could be used for the study set and test set models. The use of structural information from the target

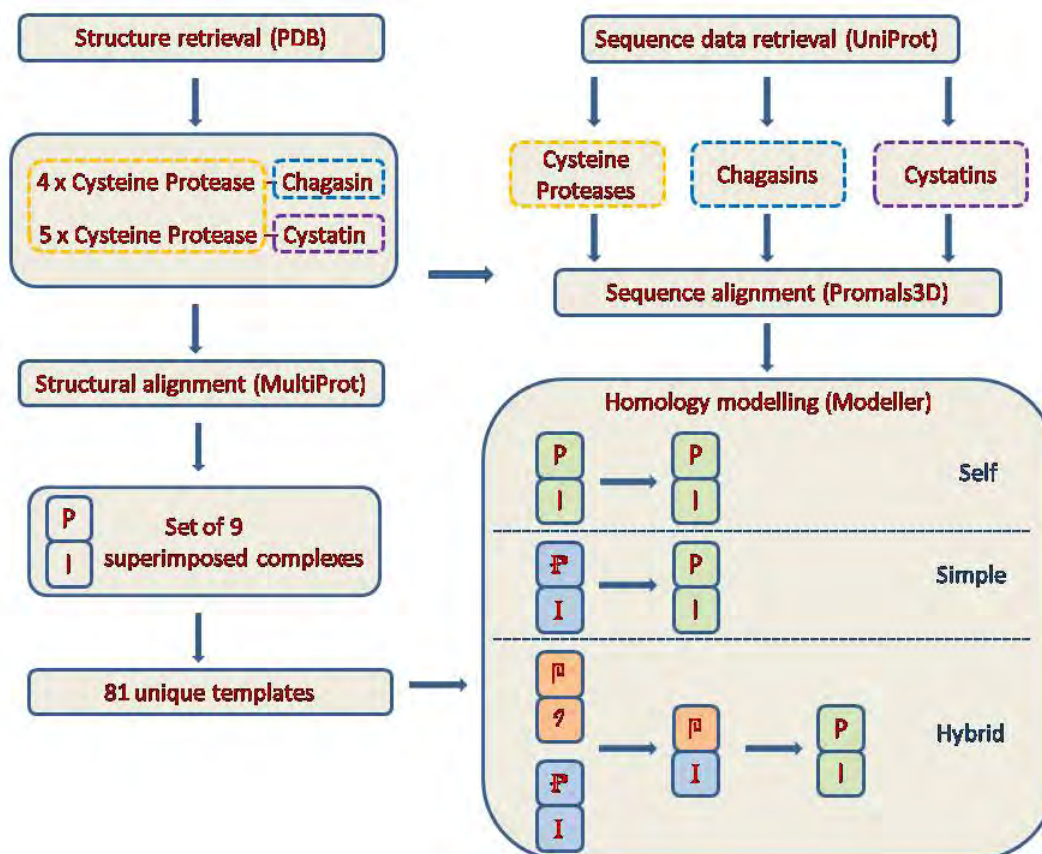


Figure 2.1 – Flow chart describing the modelling approach that was followed in the test set and the study set. Experimental complex structures were obtained from the PDB. These nine complexes were recombined using superpositioning to create 81 unique templates. Sequence data for the study proteins were obtained from UniProt. The sequence data and structure data were used together to create multiple sequence alignments for all the families in the project. The templates and alignments were then used to build models, as shown in the figure. The colors and fonts of the blocks indicate different proteins. Figure taken from Tastan Bishop & Kroon, 2011.

protein, to aid alignment, when constructing test models is questionable as this provides an unrealistically accurate alignment that would not be possible if it were a genuine modelling attempt. However, there are so many redundant structures in the dataset that one could easily be substituted for another. For this reason, the use of the same alignment for all models is justified.

2.4.3 Generation of templates

Due to results from the examination of the experimental structures (Chapter 3) and to avoid subjective decision-making, it was decided not to use any monomeric experimental structures as templates for the modelling of complexes. This left the nine, experimental, complex structures (Table 2.1) as the only source of template structures. To construct the hybrid templates, the experimental complex structures were simultaneously superimposed using MultiProt (Shatsky *et al.*, 2004). Any extra atoms in the structures were removed. This includes non-protein atoms as well multiple versions of proteins. Where multiple version of proteins were present in the experimental structures, the version with the fewest missing atoms was kept. MultiProt produces various solutions and the one with the highest number of overlapping residues that contained all the experimental structures was used. Through the recombination of the protease and inhibitor components from this set, 81 unique templates can be created (Figure 2.2). This set of 81 complex structures is referred to as the template set. Recombination is achieved by copying coordinates for the protease and inhibitor independently from the superimposed versions of the experimental structures into a single file. Nine of these templates (where protease and inhibitor are taken from the same experimental structure) are the same as the original experimental structures and the remaining seventy-two are new structures that contain information from two separate crystal structures each. The nine original crystal structures are referred to as simple templates, while those containing information from two separate experimental structures are referred to as hybrid templates. Models that were constructed using simple templates are referred to as simple models and those built on hybrid templates are referred to as hybrid models.

The following naming schema, for hybrid templates, is used in the rest of the text. A hybrid template constructed from the protease coordinates of 1YVB and the inhibitor coordinates from 3E1Z will be referred to as follows:

H:1YVB-P:3E1Z-I

The source structure of the protease is indicated with "-P" and that of the inhibitor with "-I". Simple templates are equivalent to an actual experimental structure and can thus be referred to by its actual PDB ID.

2.4.4 Template selection

The test set targets were separately modelled using all of the available templates that did not contain any coordinates from the target structure. In addition, models were also built

		1YVB	1STF	3IMA	1NB5	3K9M	2OUL	2NQD	3CBJ	3E1Z
		Ovocyst	Cyst-B	Tarocyst	Cyst-A	Cyst-A	Chag	Chag	Chag	Chag
1YVB	FP-2	S	H	H	H	H	H	H	H	H
1STF	Papain	H	S	H	H	H	H	H	H	H
3IMA	Papain	H	H	S	H	H	H	H	H	H
1NB5	Cath-H	H	H	H	S	H	H	H	H	H
3K9M	Cath-B	H	H	H	H	S	H	H	H	H
2OUL	FP-2	H	H	H	H	H	S	H	H	H
2NQD	Cath-L	H	H	H	H	H	H	S	H	H
3CBJ	Cath-B	H	H	H	H	H	H	H	S	H
3E1Z	Papain	H	H	H	H	H	H	H	H	S

Figure 2.2 – Recombination of the proteases and inhibitors of the experimental template structures yields 81 different complexes. Blue colored blocks indicate templates that contain a cystatin inhibitor and those in purple indicate templates containing chagasin. Simple templates are indicated with an S and hybrid templates with an H.

using the experimental structure itself as a template. These models are referred to as self models and serve as an estimation of the absolute maximum accuracy achievable using the modelling protocol. See Figure 2.3 for a graphical explanation of what targets were modelled using what templates. The result is that targets containing a cystatin inhibitor was modelled using 33 different templates while those containing chagasin was modelled using 25 different templates. See Figure 2.4 for the calculation of these, and other, numbers. The reason for the difference is that there are more experimental complexes containing cystatin and therefore more cystatin containing templates. The numbers shown in Figure 2.4 under "per target totals" can be obtained by counting the different kinds of available templates in Figure 2.3 and adding them up. A modelgroup refers to a set of 5 models constructed together using the same sequence and structure information (See Section 2.4.5). They differ from each other due the fact that random number generators are used in the construction of models. Since the random numbers are different for every model, different models are produced. Each member of the modelgroup thus has the same target and the same template. The combination of target and template is unique to each modelgroup in the test set (the combination may also appear in a different modelgroup in the study set).

In the study set, each target complex was only modelled based on one template. The template was taken from the template set using criteria described in Section 5.3.2.

		Cystatin containing targets									
		1YVB	1STF	3IMA	1NB5	3K9M	2OUL	2NQD	3CBJ	3E1Z	
		Ovocyst	Cyst-B	Tarocyst	Cyst-A	Cyst-A	Chag	Chag	Chag	Chag	
1YVB	FP-2	S	H	H	H	H	H	H	H	H	
1STF	Papain	H	S	H	H	H	H	H	H	H	
3IMA	Papain	H	H	X	H	H	H	H	H	H	
1NB5	Cath-H	H	H	H	S	H	H	H	H	H	
3K9M	Cath-B	H	H	H	H	S	H	H	H	H	
2OUL	FP-2	H	H	H	H	H	S	H	H	H	
2NQD	Cath-L	H	H	H	H	H	H	S	H	H	
3CBJ	Cath-B	H	H	H	H	H	H	H	S	H	
3E1Z	Papain	H	H	H	H	H	H	H	H	S	

		Chagasin containing targets									
		1YVB	1STF	3IMA	1NB5	3K9M	2OUL	2NQD	3CBJ	3E1Z	
		Ovocyst	Cyst-B	Tarocyst	Cyst-A	Cyst-A	Chag	Chag	Chag	Chag	
1YVB	FP-2	S	H	H	H	H	H	H	H	H	
1STF	Papain	H	S	H	H	H	H	H	H	H	
3IMA	Papain	H	H	S	H	H	H	H	H	H	
1NB5	Cath-H	H	H	H	S	H	H	H	H	H	
3K9M	Cath-B	H	H	H	H	S	H	H	H	H	
2OUL	FP-2	H	H	H	H	H	S	H	H	H	
2NQD	Cath-L	H	H	H	H	H	H	X	H	H	
3CBJ	Cath-B	H	H	H	H	H	H	H	S	H	
3E1Z	Papain	H	H	H	H	H	H	H	H	S	

Figure 2.3 – Templates available to cystatin and chagasin containing targets. 3IMA is given as an example containing cystatin and 2NQD as one containing chagasin. Self templates are indicated in red, with an “X”, while hybrid and simple templates are shown in blue (cystatin) or purple (chagasin). Simple templates are indicated by a darker color and an “S”, hybrid templates are indicated with an “H”. Templates shown in gray were not used to model the particular target. Cystatin containing targets are 1YVB, 1STF, 1NB5, 3IMA and 3K9M. Chagasin containing targets are 2OUL, 2NQD, 3CBJ and 3E1Z.

Targets		Modelgroups		Models
Project total:	119	375	1875	
Test set	9	265	1325	
	5	Containing cystatin	165	825
	Per target totals		33	165
		Self	1	5
		Simple	4	20
		Hybrid	28	140
	4	Containing chagasin	100	500
	Per target totals		25	125
		Self	1	5
		Simple	3	15
		Hybrid	21	105
Study set	110	110	550	

Figure 2.4 – Total number of models built. The “per target totals” is multiplied by the respective number of targets to give the totals for cystatin and chagasin containing targets respectively.

2.4.5 Modelling protocol

Given a template and target complex, the correct alignments of proteases and inhibitors are automatically extracted from the multiple sequence alignments and used to generate a PIR alignment (file format required by Modeller) and a modelling script using custom software (see Section 2.5). The modelling script, PIR alignment and template structure, in PDB format, is then submitted to Modeller 9.6 which produces 5 different models, as specified in the script. These 5 models are referred to as a modelgroup. The reason for building 5 models instead of a different number is that it provides enough models to get significant results while maintaining computational time within practical limits. It is not only the computational time of the modelling itself that has to be considered, but also that of any further processing of the models. The default automodel routine was used with refinement set to very slow. The modelling scripts and PIR alignments are available on the website accompanying the thesis (<http://rubi.ru.ac.za/cpmdb>).

2.4.6 Calculation of scoring functions

Two different scoring functions, Rosetta energy (www.rosettacommons.org) and DOPE Z-score (Sali & Shen, 2006), were calculated for all models produced, before and after energy minimization. Before minimization, Rosetta energy was calculated using the "score" program of the Rosetta 3.1 package with the included database. The only change to the default arguments was setting "no_optH" to false. After minimization, the Rosetta energy is written to a file as part of the standard procedure of the minimization protocol. Thus the score was simply read from the file. The DOPE Z-score was calculated with the default parameters using the "assess_normalized_dope" function. It was calculated in the same way before, and after, energy minimization.

2.4.7 Energy minimization

Energy minimization was carried out using the "relax" protocol of the Rosetta 3.1 package, using the included database. Default parameters were used. The "relax" protocol produces an all-atom structure, including hydrogens, by applying aggressive Monte Carlo sampling to find conformations that minimize the Rosetta energy. The Rosetta source code and database was downloaded from www.rosettacommons.org.

2.5 Information management

Throughout the project, programs and classes were written in the python programming language to automate processes and thereby accommodate the high-throughput nature of the

study. Object orientated design (software design pattern) was used and classes were created to represent PDB structures, alignments and modelgroups amongst others. A complete description of all the software written is beyond the scope of this document. The programs allowed the programmatic manipulation of all the data. For the analysis of the results, the data was serialized into a comma separated value file and loaded into R (<http://r-project.org>). R is a powerful program designed for statistical analyses. It has powerful data-manipulation and graphing functionality that was not afforded by the programs that were used in the modelling process. During the creation of the website, tables were created in MySQL(<http://www.mysql.com>) to accommodate some of the classes in a database form, for fast access by the webserver.

2.6 Computational resources

The project required approximately seven thousand of hours of CPU computation. The energy minimization step in particular takes up a lot of computational time. Various computational resources were utilized. A clusters of several computers were made available by the Chemistry Department at Rhodes University. The main bioinformatics server at Rhodes University and several spare computers in the bioinformatics laboratory were also used. Lastly, toward the end of the project about twenty computers in a laboratory at the Bioinformatics and Computational Biology Unit (BCBU) at University of Pretoria were made available. Because of the diversity of computing resources, a modular job distribution system was created. The job distribution contains two programs, the distributor and the worker. The distributor runs on the central computer that holds all the project data and copies the necessary files to the computers that actually run the jobs. The worker program accepts the files, carries out the jobs and copies the results back to the central computer. Because of different system and network configurations, small changes were necessary to the worker program in each case. In the case of the cluster, which had job distribution software installed, the worker program was modified to submit the job to the queue instead of carrying out the jobs itself.

Chapter 3

Examination of crystal structures

3.1 Chapter overview

This chapter describes an investigation of several experimental structures that are important to this project. The chapter begins with discussion of the aim and motivation for doing this investigation. The methods used are then described and the structures that were investigated are listed. A short justification of how the method is connected to the achievement of the aims is also given in the method section. The results are then given, describing differences between the sub-families and pointing out significant conformational changes where present.

3.2 Aim and motivation

Before setting out on a high-throughput modelling project, it is prudent to first examine the available experimental structures of the protein families involved. This will yield valuable information for informed decision-making in template selection and provide context for the analysis of the results. Two important questions were addressed in this part of the project.

The first question is whether the proteins undergo any conformational changes upon binding. If there is a conformational change, then the template must be of the same state as the target protein, i.e. for modelling complexes, the template must be part of a complex structure. If there are conformational changes, do they vary for different binding partners? If so, then templates can only be taken from complexes with the same class of binding partner. If not, templates can be taken from any complex.

The second important question is that of what constitutes an accurate model. How accurately do crystal structures represent the native protein? How much deviation from the crystal structure can be present in a model and still be regarded as having accuracy comparable to that of a crystal structure? This is an important question as the accuracy of the models must be measured against some standard. At a certain threshold, closer similarity to the crystal structure

is not indicative of a more native-like structure. It is important to have an idea of what that threshold is so that models can be judged within that context.

Comparisons of the different sub-families were also carried out to give some information on the feasibility of cross-family modelling and simply to provide some additional background information.

3.3 Method

To investigate conformational change upon binding, experimental structures of the same protein in bound and unbound states were compared to each other. This was done for all of the protein families involved in the project. Experimental structures to be compared were identified in the PDB through manual searching. As a starting point, the experimental template structures (Table 2.1) were used. These structures represent all complex structures of C1 proteases and protein inhibitors available. As such, any protein with both monomeric and complex experimental structures must be present in this group. The experimental structures of every protein in that group were identified using BLAST searching. From this information, a representative protein in each sub-family was chosen based on the number of different structures available. Table 3.1 shows all the experimental structures that were used in this chapter. All the structures of each protein were compared to each other. Structures were compared using visual inspection by means of molecular visualization software as well as calculation of RMSD values.

For the question, as to what constitutes an accurate model, multiple versions of cathepsin-B in the same experimental structure (PDB ID: 1GMY) were compared to each other to estimate the accuracy of the experimental structure itself. Some background information is required to understand the rationale behind this:

When proteins are crystallized for structure determination, there are sometimes more than one protein or protein complex per unit cell (smallest repetitive unit) of the crystal. When the structure is then solved, the result is multiple structures of the protein or complex being investigated. There are always some minor differences between these structures. The difference between any experimental structure and the real, native, structure of a protein can be thought of as having three different components. The first, and unavoidable, component exists because real proteins are mobile, and this mobility can not be represented in the structure formats (like the PDB file format) used to present protein structures. The second component is due to conformational changes in the protein induced by experimental manipulation of the protein. The third component is the experimental error and limitation in determining the structure of the protein as it exists in the crystal (when using X-ray diffraction) or solution (when using NMR). None of these components are dependent on the number of proteins in the crystal unit cell and it is therefore reasonable to assume that these errors are reflected in the variation between different proteins structures determined from the same crystal. The maximum theoretical accuracy of homology models, as measured by comparison to experimental structures, can be no higher than the accuracy of the experimental structure itself. In

Table 3.1 – Experimental structures examined in structural investigation. Shown is PDB ID, experimental technique as well as name, species and sub-family of protease and inhibitor.

PDB ID	Exp. tech	Protease	Organism	Sub-family	Inhibitor	Organism	Sub-family
Cathepsin-L like cysteine proteases - papain							
1PPN	X-ray	Papain	<i>C. papaya</i>	Cathepsin-L like			
1STF	X-ray	Papain	<i>C. papaya</i>	Cathepsin-L like	Cystatin-B	<i>H. sapiens</i>	Type I
3E1Z	X-ray	Papain	<i>C. papaya</i>	Cathepsin-L like	Chagasin	<i>T. cruzi</i>	Chag
3IMA	X-ray	Papain	<i>C. papaya</i>	Cathepsin-L like	Taro-cystatin	<i>C. esculenta</i>	Type II
Cathepsin-B like cysteine proteases - cathepsin-B							
1GMY	X-ray	Cathepsin-B	<i>H. sapiens</i>	Cathepsin-B like			
3CBJ	X-ray	Cathepsin-B	<i>H. sapiens</i>	Cathepsin-B like	Chagasin	<i>T. cruzi</i>	Chag
3K9M	X-ray	Cathepsin-B	<i>H. sapiens</i>	Cathepsin-B like	Cystatin-A	<i>H. sapiens</i>	Type I
Type I cystatins - cystatin-A							
1DVC	NMR				Cystatin-A	<i>H. sapiens</i>	Type I
1NB5	X-ray	Cathepsin-H	<i>S. scrofa</i>	Cathepsin-L like	Cystatin-A	<i>H. sapiens</i>	Type I
3K9M	X-ray	Cathepsin-B	<i>H. sapiens</i>	Cathepsin-B like	Cystatin-A	<i>H. sapiens</i>	Type I
Type II cystatins - ovocystatin							
1CEW	X-ray				Ovocystatin	<i>G. gallus</i>	Type II
1A67	NMR				Ovocystatin	<i>G. gallus</i>	Type II
1YVB	X-ray	Falcipain-2	<i>P. falciparum</i>	Cathepsin-L like	Ovocystatin	<i>G. gallus</i>	Type II
Chagasin-like inhibitors - chagasin							
2H7W	X-ray				Chagasin	<i>T. cruzi</i>	Chag
2NQD	X-ray	Cathepsin-L	<i>H. sapiens</i>	Cathepsin-L like	Chagasin	<i>T. cruzi</i>	Chag
2OUL	X-ray	Falcipain-2	<i>P. falciparum</i>	Cathepsin-L like	Chagasin	<i>T. cruzi</i>	Chag
3E1Z	X-ray	Papain	<i>C. papaya</i>	Cathepsin-L like	Chagasin	<i>T. cruzi</i>	Chag
3CBJ	X-ray	Cathepsin-B	<i>H. sapiens</i>	Cathepsin-B like	Chagasin	<i>T. cruzi</i>	Chag

other words, when the variation between a model and the crystal structure of a protein is of the same magnitude as between different conformations in the same experimental structure, the model can be regarded as having equal accuracy to the experimental structure.

RMSD values were calculated using the McLachlan algorithm (McLachlan, 1982) as implemented in the program ProFit (authored by Martin, A.C.R. and Porter, C.T, downloaded from <http://www.bioinf.org.uk/software/profit/>) unless stated otherwise. All backbone atoms were used in the calculation of RMSD values. Figures were created using PyMOL (DeLano, 2002). The superpositioning shown in the figures (superpositioning performed by PyMOL algorithm) is not always the same as that used for calculating the RMSD values. The Protein Interaction Calculator (PIC) was used to determine which residues formed part of the interface (Tina *et al.* , 2007). A residue was deemed part of the interface if it shows an interaction with the other protein in any of the available experimental structures. While this method is consistent, it leads to unequal interfaces between type I and type II cystatins because type I cystatins have more experimental complex structures available.

3.4 Results and discussion: Papain-like cysteine proteases

3.4.1 Comparison between cathepsin-L like and cathepsin-B like sub-families

Cathepsin-B and cathepsin-L were chosen for this comparison as they are the defining proteins of the two sub-families. Figure 3.1 shows an overview of the structural differences and similarities between the two proteins. The biggest structural difference between the two sub-families is the presence of the occluding loop (shown in green, on the left) in members of the cathepsin-B like subfamily. This insert is the defining difference between the subfamilies. Apart from the occluding loop there is also a conformational difference on the opposite side of the catalytic residues. The loop indicated in green in cathepsin-B on the right of the figure is much closer to the active site cleft than in cathepsin-L. Given these differences, it would be unwise to build models using a template from a different sub-family than the target protein.

3.4.2 Cathepsin-L like subfamily

Papain (*C. papaya*) was chosen to represent the cathepsin-L subfamily because it is the oldest member and has the most experimental structures available. The structures chosen for comparison can be viewed in Table 3.1.

Figure 3.2 shows the overall structure of the protein and it can be seen that there are negligible conformational differences between the backbone conformations in the various crystal structures. The side-chains also exhibit very little change and the side-chains that do differ (TYR 61, TYR 67 and TRP 69, using sequential numbering) are solvent exposed. This means

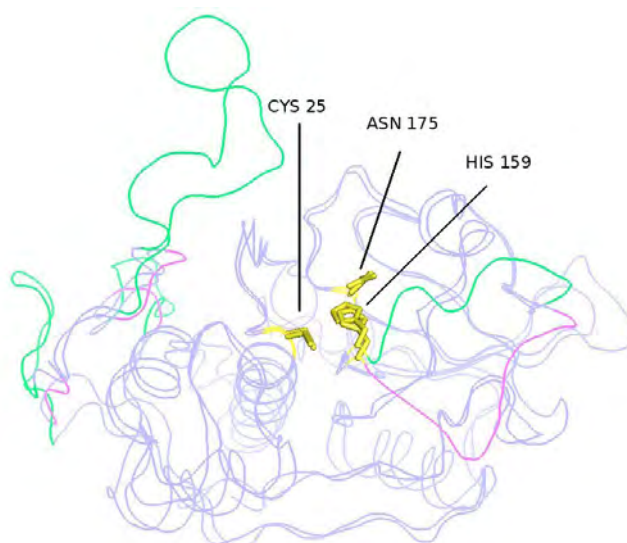


Figure 3.1 – A visual comparison of the cathepsin-L like and cathepsin-B like subfamilies. The proteins shown are cathepsin-B (PDB ID: 3K9M) and cathepsin-L (PDB ID: 2NQD). Where the proteins are the same, the backbone is colored blue. Where the proteins differ, cathepsin-B is indicated in green and cathepsin-L in pink. The catalytic triad residues are shown in yellow.

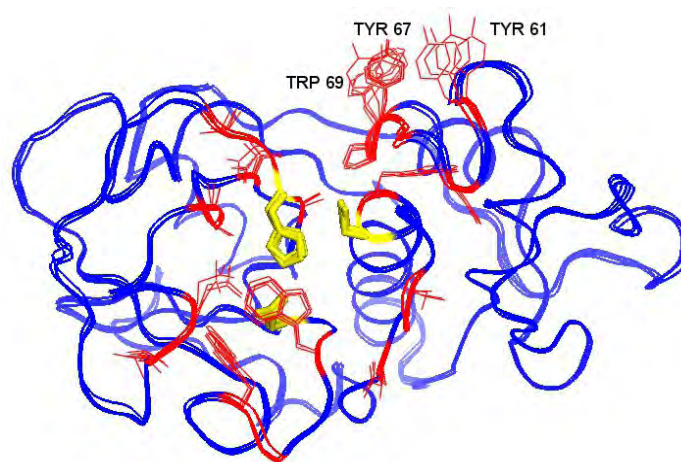


Figure 3.2 – Four experimental structures of papain (1PPN, 1STF, 3E1Z, 3IMA) were superimposed and shown in the ribbon representation. The residues that contact the protein inhibitors are shown in red with their side-chains shown in line presentation. The catalytic triad residues are shown in yellow in stick presentation.

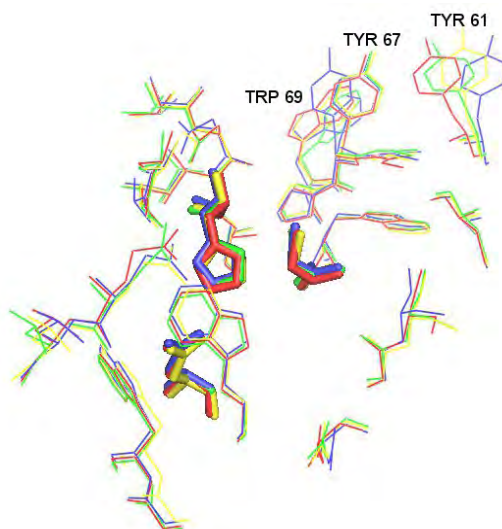


Figure 3.3 – The interacting residues of 1PPN(red), 1STF(green), 3E1Z(blue) and 3IMA(yellow) are shown in line representation and the catalytic triad in stick representation.

Table 3.2 – Backbone RMSD in Å between selected structures of papain.

	1STF	3E1Z	3IMA
1PPN	0.383	0.362	0.347
1STF		0.396	0.413
3E1Z			0.382

that for modelling purposes it will not make a big difference which of the crystal structures are used as a template.

Figure 3.3 shows a close up of the interacting residues. The structure that differs most is 3E1Z with two side-chain changes (TYR 67 and TYR 61). In fact, two positions are specified for TYR 67 in the 3E1Z experimental structure, one that is visible slightly to the left compared with the other structures and one overlapping with them. With TYR 61, all the structures show some variation so it is likely mobile. Thus, the differences between the four structures are almost negligible.

Table 3.2 shows the RMSD values between the different experimental structures of papain used. It shows that the most different conformations, as measured by RMSD, are those from 1STF (Cystatin-B complex) and 3IMA (Tarocystatin complex) with a value of 0.413 Å. This is consistent with no conformational change upon binding as the monomeric conformation of 1PPN is no more different from the complex structures as they are from each other. The 3E1Z(Chagasin) also does not show an increased RMSD when compared to the two cystatin complexes, supporting the hypothesis that there is no conformational change when binding to a cystatin or chagasin-like inhibitor.

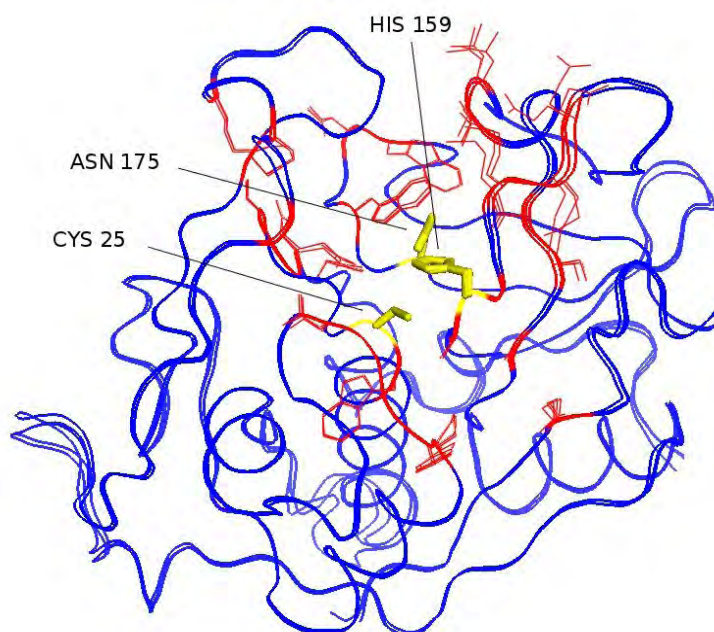


Figure 3.4 – The different conformations of cathepsin-B in 1GMY were superimposed and shown in the ribbon representation. The side-chains of the residues that contact the protein inhibitors are shown in red with their side-chains shown in line presentation. The catalytic triad residues are shown in yellow, in stick representation.

3.4.3 Cathepsin-B like subfamily

Human Cathepsin-B was chosen to represent the cathepsin-B like subfamily. It is the only member of the subfamily with experimental structures of complexed forms. The structures chosen for comparison can be viewed in Table 3.1. There is also a monomeric experimental structure of cathepsin-B available (1GMY) with three different conformations in the same unit cell of the crystal.

Different conformations of the monomeric structure in 1GMY gave the opportunity to assess RMSD variation under the same conditions of crystallization. This addresses the second aim of this chapter as explained in Section 3.2.

Figure 3.4 shows the overall structural variation between different versions of the cathepsin-B structure within the crystal structure 1GMY; very little is visible, as expected. If we compare this to Figure 3.2 we see that the amount of variation is almost the same. It was also found that the occluding loop didn't show any more variation than the rest of the protein although this isn't highly visible in the figure. The implication of this is that while the occluding loop is flexible, there is a clearly favoured conformation while in monomeric form.

Table 3.3 shows the RMSD between the different conformation of the cathepsin-B structure within the crystal structure 1GMY. It is actually slightly more than between the different conformations of papain shown in Table 3.2. This shows that the variation between the various

Table 3.3 – Backbone RMSD in Å between different cathepsin-B conformations of 1GMY.

	B	C
A	0.508	0.386
B		0.566

papain conformations is very small, but also that experimental error varies from one experimental structure to the next. This deduction is made from the observation that the variation between the papain conformations, from different experimental structures, is lower than the variation between the cathepsin-B conformations currently under discussion. This implies that the papain experimental structures are either more accurate or that they all have correlating errors. As mobility of proteins vary from protein to protein and mobility is a large contributor to errors in experimental structures, it makes sense that the accuracy of experimental structures should also vary. Taking this into consideration, a value of 0.38Å (median from Table 3.2 and minimum in table 3.3) will be considered a rough lower limit of the RMSD a homology model can achieve.

Figure 3.5 shows cathepsin-B in different conformations from three experimental structures including one monomeric structure and two complex structures. Only the first conformation from 1GMY was used for this comparison. This relates to the aim of determining if conformational change takes place upon binding. Conformational change can be clearly seen in the occluding loop when interacting with protein inhibitors. As such we cannot expect to model the occluding loop correctly if we don't have a template with the correct loop conformation. Unfortunately there are not enough experimental structures available to determine how much the movements varies between different inhibitors. The movement is clearly different between chagasin (3CBJ) and cystatin-A (3K9M); it would be informative to have a structure with a type II cystatin, to compare the change that takes place. However, no such experimental structure is available. This implies difficulty in the modelling of cathepsin-B complexes.

Figure 3.6 shows a close up of the interface residues and more clearly illustrates the differences between the individual conformations of the occluding loop. In the monomeric conformation of 1GMY, shown in blue, there are some residues that lie within the active site cleft but are pushed away in both the complexed conformations to make space for the inhibitors. It can also be seen that it is only the residues of the occluding loop (residue 109 to 126) that undergo conformational change. The figure clearly shows that all three structures are unique and the change induced upon binding to chagasin is different from the change upon binding to cystatin. The implication for modelling is that there will be uncertainty of the occluding loop conformation in any model, until more experimental structures of cathepsin-B in complex with protein inhibitors are published.

Table 3.4 shows the RMSD between the different cathepsin-B structures. It confirms that there is conformational change. It also indicates that the change is different in the case of binding to cystatin-A from the change in the case of binding to chagasin. Based purely

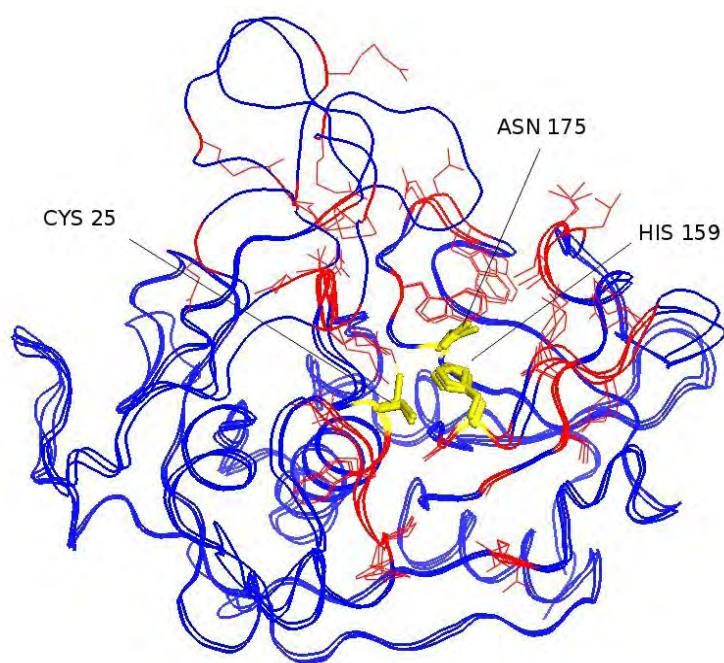


Figure 3.5 – Three experimental structures of cathepsin-B (1GMV, 3K9M, 3CBJ) were superimposed and shown in the ribbon representation. The side-chains of the residues that contact the protein inhibitors are shown in red with their side-chains shown in line presentation. The catalytic triad is shown in yellow in stick presentation. At the top of the figure it can be seen that there are big conformational differences between the various structures.

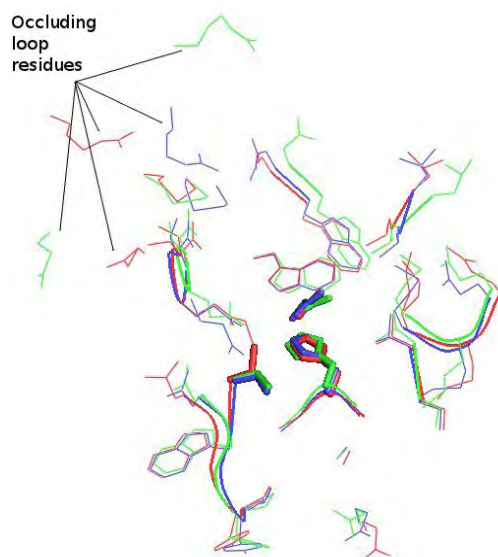


Figure 3.6 – The interacting residues of 1GMV(blue), 3K9M(green) and 3CBJ(red) is shown in line representation and the catalytic triad residues in stick representation.

Table 3.4 – Backbone RMSD in Å between selected structures of cathepsin-B.

	3K9M	3CBJ
1GMV	2.759	2.101
3K9M		2.275

on the RMSD values it seems that the conformational change is bigger for cystatin-A than for chagasin although this cannot be used to infer anything about the affinity between the inhibitors and the protease.

3.5 Results and discussion: Cystatin superfamily

3.5.1 Comparison between type I and type II cystatins

Cystatin-B, ovocystatin and tarocystatin was chosen for this comparison. Two type II cystatins are included to highlight a slight difference between them. Figure 3.7 shows a global comparison of the structures of the type I and type II cystatins. There are several differences between the type I cystatins and the type II cystatins.

The main structural differences are the presence of a C-terminal extension in type I cystatins, a bend in the α -helix of type I cystatins and an insert between the second and third β -strand in type II cystatins. Two type II cystatins are given to show the difference between phytocystatins (tarocystatin, 3IMA) and other type II cystatins (ovocystatin, 1YVB). In phytocystatins the insert in the β -sheet is missing. Tarocystatin therefore has characteristics of both sub-families. Since it shares more characteristics and higher sequence identity with the type II cystatins it is classified as a type II cystatin. The differences mentioned are all visible on the figure. The type I C-terminal extension is visible in pink on the right of the figure. The α -helix bend can be seen on the left where the pink and green helices differ. The insert between the second and third strands of the β -sheet is visible at the top where the type I loop is visible in pink, the phytocystatin loop in dark-green and the type II loop in lime-green. An important sequence difference is the presence of a tryptophan residue (residue 80, 3IMA; residue 99, 1YVB) in the C-terminal-most binding loop of type II cystatins (bottom left of the figure).

Of the four differences between the groups there are two that occur at the interface with the protease. These are the tryptophan residue (with a very large hydrophobic sidechain) on the C-terminal-most binding loop and the C-terminal extension. It is not shown in this figure but the tryptophan side-chain forms extensive hydrophobic interactions with the protease during inhibition. The C-terminal extension of cystatin-B in 1STF interacts with the protease during inhibition and thus the type I cystatins actually have four contact points with the protease, in contrast the three mentioned in section 1.4.4.

Modelling of type I cystatins using type II templates would lead to incorrect modelling of the C-terminal extension while the loop on the non-interface side in type II cystatins would

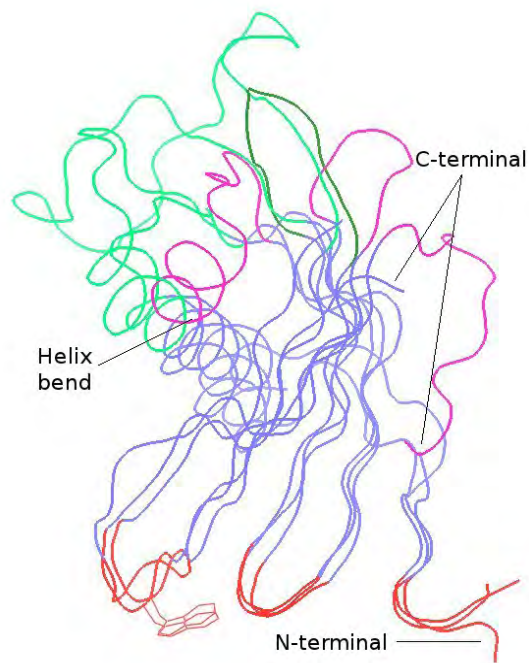


Figure 3.7 – The figure shows ovocystatin (1YVB), tarocystatin (3IMA) and human cystatin-B (1STF). Where the proteins are the same, they are colored in blue. Where the proteins differ, cystatin-B is colored pink while ovocystatin and tarocystatin are colored lime-green. Where tarocystatin deviates from the normal type II pattern it is colored in dark-green. The binding loops are colored in red.

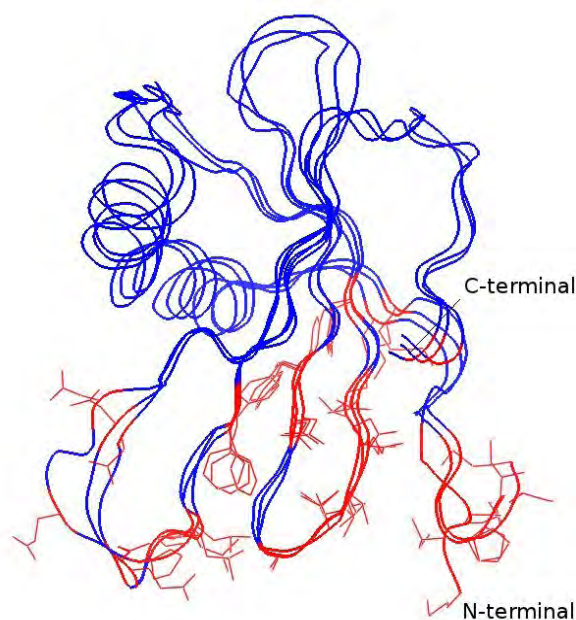


Figure 3.8 – Three structures of cystatin-A (1DVC, 1NB5, 3K9M) were superimposed and shown in the ribbon representation. The side-chains of the residues that contact the protease are shown in line presentation in red.

be incorrectly modelled using type I templates. Whether a modelling program would be able to account for the helix difference is doubtful and cross sub-family modelling would likely produce an incorrect helix conformation as well.

3.5.2 Type I cystatins

Cystatin-A was chosen to present this family as there are two experimental structures available in protein complexes. Unfortunately only an NMR structure is available for the monomeric conformation. The structures used can be viewed in Table 3.1. The single NMR structure (1DVC) was produced by performing energy minimization on an averaged structure of the NMR ensemble by the original publishers of the structure.

Figure 3.8 shows an overview of the three cystatin-A structures. The whole protein is somewhat flexible and shows more variation than the proteases, apart from the cathepsin-B occluding loop. Some of this difference is due to the use of an NMR structure and comparing it to crystal structures. There is a big difference in the N-terminal (bottom right) conformation between the monomeric structure and the complexed forms. There is even a slight difference in the C-terminal-most loop (bottom left) between the monomeric and two complexed forms. The implications for modelling is that higher RMSD scores than for the proteases can be expected as the protein is more mobile. It is also likely that using monomeric templates for complex structures would lead to significant errors in the interface, especially in

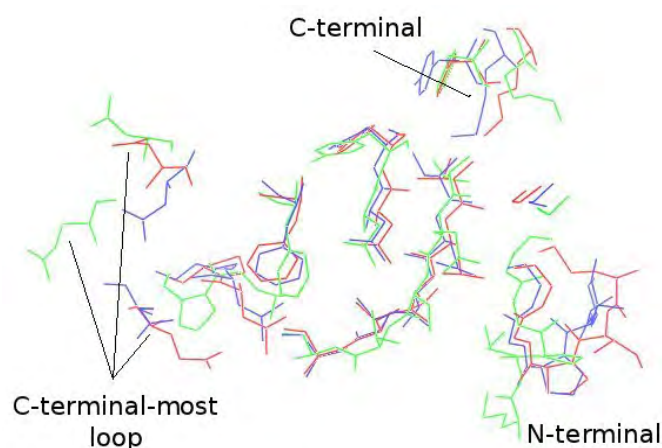


Figure 3.9 – The interacting residues of three cystatin-A structures, 1DVC(green), 1NB5(red) and 3K9M(blue) is shown in line representation.

Table 3.5 – Backbone RMSD in Å values between indicated structures of cystatin-A.

	1NB5	3K9M
1DVC	2.005	1.999
1NB5		1.058

the N-terminal.

Figure 3.9 shows the interacting residues of the different cystatin-A structures. The N-terminal is on the right and the C-terminal-most loop is on the left. The big conformational changes upon binding are more clearly visible than on the previous figure. Especially the C-terminal-most loop and the N-terminal is completely different between the monomeric and complex structures. Small differences are also visible between the cathepsin-H(1NB5) and the cathepsin-B(3K9M) complex structures. These differences are mostly restricted to side-chain conformational changes and models of one based on the other could be fairly accurate. The biggest difference between the backbone conformations of the two complex structures is just above the C-terminal-most loop (visible in the top-left) where the backbones differ and the side-chains point in opposite directions.

Table 3.5 shows the RMSD between the different experimental structures of cystatin-B. The biggest difference is between the monomeric structure and the complex structures. There is also a significant difference between the conformations of the two complex structures.

For modelling, using monomeric structures as templates for complex models would lead to significant errors. Using templates from complex structures where the protease is of the other sub-family than in the target may introduce slight errors but should achieve good overall accuracy.

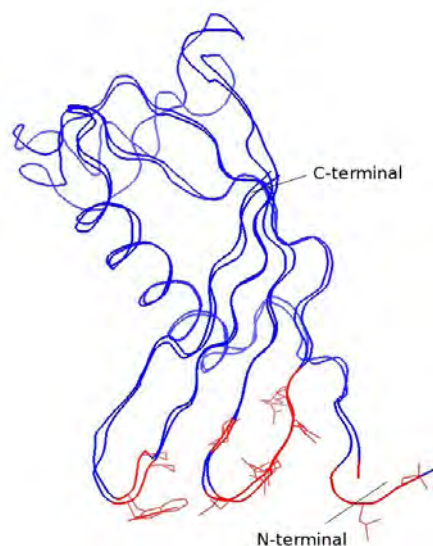


Figure 3.10 – Two structures of ovocystatin (1CEW, 1YVB) were superimposed and shown in the ribbon representation. The side-chains of the residues that interact with the protease are shown in red with their side-chains shown in line presentation.

3.5.3 Type II cystatins

Ovocystatin was chosen to represent this subfamily as it is the only member of the family to have been solved in complex with a protease that also has a monomeric structure available. Unfortunately there is no complex structure with a cathepsin-B like protease. The structures used for the comparison can be viewed in Table 3.1.

Figure 3.10 shows a global overview of the ovocystatin structures. There is substantial flexibility in the loops, at the end of the molecule, furthest away from the binding site (visible at the top of the figure). The tryptophan residue, that is characteristic of the type II cystatins, is clearly visible at the bottom-left of the picture. The side-chain points away from the core of the cystatin, even in the monomeric form. In the monomeric crystal structure the N-terminal was not resolved. Many studies have indicated that the N-terminal increases the binding affinity and that truncation of the N-terminal reduces the affinity significantly Hall *et al.* (1995). When binding to a protease the N-terminal becomes anchored and static, in the monomer it is mobile. The other two loops show similar conformations between the monomeric and complexed form. Using the monomeric structure as template for complex models would lead to incorrect modelling of the N-terminal, a very important part of the interaction.

Figure 3.11 shows the interacting residues of ovocystatin. Differences are visible in the C-terminal most loop (on the left). The orientation of the residues are very similar but they show translational movement relative to each other. The middle loop is very similar but does show some variation in side-chain conformations. The main difference in the N-terminal is the number of residues visible in the experimental structures. Three residues are missing in

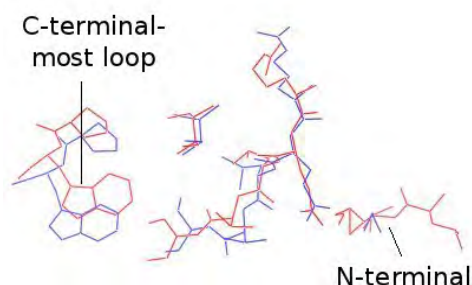


Figure 3.11 – The interacting residues ovocystatin in 1CEW(blue) and 1YVB(red) are shown in line representation.

the monomeric structure. Given the importance of the N-terminal region of cystatins, the modelling of complexes using monomeric templates should be avoided.

The RMSD between 1YVB and 1CEW is 2.724Å, and this does not include the unresolved N-terminal region. This is a substantial amount of variation and backs up the findings of the visual inspection. The variation leading to the high RMSD is mainly in the loops, away from the interface. The variation in the interface area is much less.

The PDB structure 1A67 was also investigated. 1A67 is a monomeric NMR structure containing ovocystatin. The various solution structures published under 1A67 is shown together with the X-ray structure 1CEW in Figure 3.12. This gives an indication of the natural flexibility of the protein and hence is relevant to the question of what accuracy can realistically be expected of a model.

The NMR data shows that the β -sheet part of the molecule is stable. There is a lot of flexibility in the loop area (visible at the top of the figure), away from the interacting site. There is a significant difference in the distance between the α -helix and β -sheet when comparing the NMR structures to the crystal structure (not visible in figure). Since the loop at the top is so flexible, it follows that it cannot be correctly represented in a single PDB structure and thus cannot be modelled accurately using a structure format where each atom is given a single position like the PDB file format.

The NMR structures were superimposed on the first one and backbone RMSD then calculated relative to the average coordinates using VMD

(Visual Molecular Dynamics <http://www.ks.uiuc.edu/Research/vmd/>). The results are shown in table 3.6. They vary from 1.517Å to 2.790Å. This is indicative of the flexibility of the monomeric protein in solution and reinforces the points made above.

3.6 Results and discussion: Chagasin-like inhibitors

Chagasin itself was chosen to present this family as there are plenty of complex structures and a monomeric crystal structure available. The structures used for comparison are indicated in

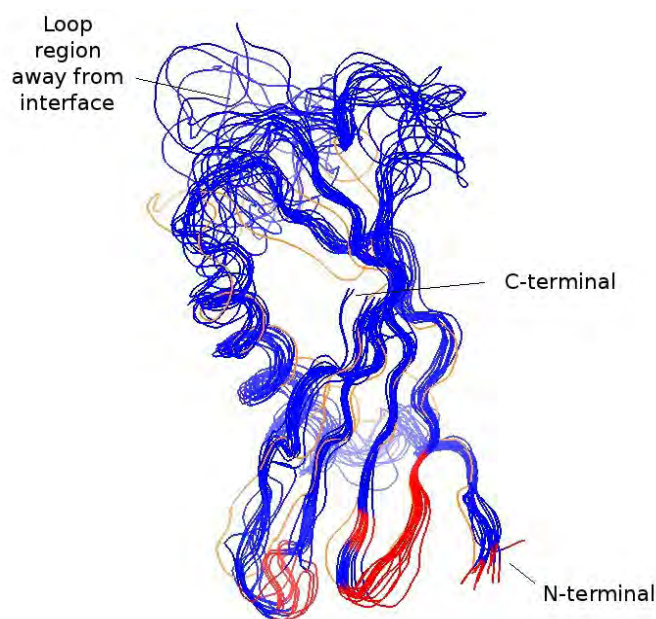


Figure 3.12 – Ovocystatin from 1CEW (crystal structure) in orange and from 1A67 (NMR solution structures) in red (interface) and blue.

Table 3.6 – Backbone RMSD in Å between NMR solution structures of 1A67 and their averaged structure.

	RMSD
1	1.517
2	2.054
3	2.117
4	2.139
5	2.790
6	2.417
7	2.512
8	2.258
9	2.100
10	2.450
11	1.728
12	1.877
13	1.756
14	1.634
15	2.384
16	2.156

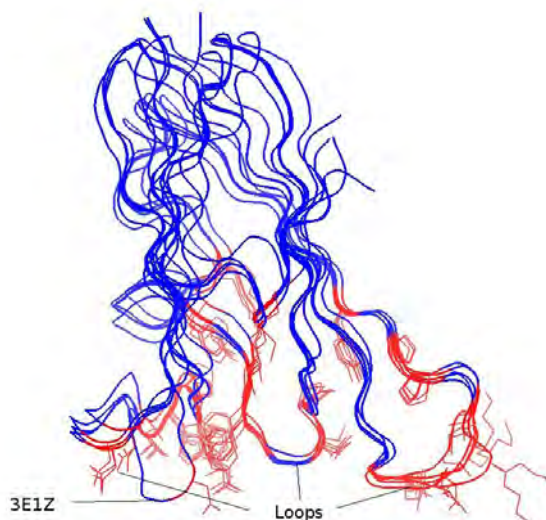


Figure 3.13 – Five structures of chagasin (3H7W, 2OUL, 3CBJ, 2NQD and 3E1Z) were superimposed and shown in ribbon representation in blue. Interface residues are shown in red with their side-chains in line representation.

Table 3.1.

Figure 3.13 shows an overall view of the different chagasin structures. The middle binding loop clearly has the most consistent conformation while there is substantial variation in the outer loops. The body of the protein, away from the interface also show substantial variation. In the loop on the right, the variation is due to side-chain as well as backbone conformational change. The loop on the left (in red) shows little conformational change except in 3E1Z in which a residue from that loop interacts with papain. That interaction with papain seems to induce a conformational change not present in any of the other complexes, nor in the monomer.

Figure 3.14 shows the interacting residues of the different chagasin structures. The big change in 3E1Z relative to the others can be clearly seen on the far left. The pink aspartic acid residue (number 99 in original 3E1Z numbering) is on its own, far away from the others. Overall there does seem to be some variability in the whole interface region. It must be noted that the flexibility of the protein influences the superpositioning of interface residues which could exaggerate the apparent variation.

The RMSD values in Table 3.7 show the variation between the individual chagasin structures. The variation between the monomer and the complexes is of the same order as between the complexes themselves. This indicates little conformational change upon binding and significant flexibility in all forms. The most unique chagasin conformation is that of the complex with papain.

When modelling a chagasin complex, use of the monomeric conformation as template could possibly lead to an accurate model but has no advantages over using a complex conformation.



Figure 3.14 – The interacting residues if 2H7W(red), 2NQD(green), 2OUL(blue), 3CBJ(yellow) and 3E1Z(pink) are shown in stick representation.

Table 3.7 – RMSD values of indicated chagasin structures.

	2NQD	2OUL	3CBJ	3E1Z
2H7W	0.826	0.643	0.964	1.083
2NQD		0.531	0.733	1.127
2OUL			0.762	1.062
3CBJ				1.204

Using the conformation from 3E1Z is not advised as it shows a unique interaction.

3.7 Conclusion

The goal of this chapter was to gather information relevant to modelling decisions from the experimental structures and provide context to the rest of the project. Also, to provide an idea of what can be viewed as an accurate model. This was done by comparing different crystal structures of individual proteins to each other. A comparison like this is limited by the availability of experimental structures. Ideally one would have experimental structures of the bound and unbound forms of the same protein available. Having multiple complex structures with different binding partners is also helpful. More experimental structures would have been useful, especially of type II cystatins. An experimental structure of cathepsin-B bound to ovocystatin would be very interesting.

The findings showed that papain and chagasin do not undergo conformational change upon binding. However, in the cases of the cathepsin-B, cystatin-B and ovocystatin, conformational change were detected. As a result, the modelling of complexes using monomeric experimental structures should be avoided. Only in the case of cathepsin-L like proteases can the use of templates derived from monomeric structures be beneficial. In the case of chagasin, using the monomeric structure as a template is viable but does not offer any advantages over using the complex structures. Unlike with cathepsin-L like proteases, there are chagasin-like inhibitors with experimentally determined monomeric structures that do not also have an experimental complex structure.

It was found that the accuracy of the experimental structures examined vary from about 0.38Å RMSD upwards. The mobility/flexibility of a protein has a big influence on the RMSD value that should be viewed as accurate. The NMR structures differed remarkably from the X-ray diffraction structures, showing that the experimental technique used to produce experimental protein structures has a substantial impact on the structure produced. Hence, models produced through the use of computer based structure prediction may be accurate, even if they are not exactly the same as the experimental structures. In theory, the application of computer modelling techniques may actually improve the accuracy of experimentally produced protein structures. In the case of X-ray diffraction, the structures do not represent the actual protein structure under physiological conditions. With NMR, the problem is that the technique does not uniquely describe the structure but rather an ensemble. Some of the members of this ensemble may not actually occur in nature but are simply indistinguishable from the native structure through the data recorded. In both cases, theoretical physics and chemistry consideration could be used to improve the structure and these can only be applied through the use of computers, due to the computationally intensive nature of the problem. A major obstacle is that there is no way to confirm whether theoretical predictions that go beyond the accuracy of experimental techniques is actually correct.

Chapter 4

Construction and evaluation of test models

4.1 Chapter overview

This chapter describes the high-throughput modelling, evaluation and analysis of the test set as described in Section 2.3.2. The chapter starts with a discussion of the aim and motivation for carrying out such an extensive evaluation of modelling results. The method followed is then described. The results and discussion are then given in the form of a summary and an analysis of the produced data. Specific questions are addressed by graphing and calculations of statistical metrics. A conclusion is given at the end.

4.2 Aim and motivation

The work of this chapter addresses some of the questions posed in Section 2.2. It aims to answer the question of how to best choose a template for the modelling of C1 protease complexes from the many possibilities available. This is a very important question as the selection of the correct template has a major influence on the accuracy of any homology model. The chapter also addresses the question of whether hybrid templates can produce similar quality models to simple templates. This can be seen as an extension of the previous question as hybrid templates are simply a subset of available templates. The test set models, built in this chapter, also gave the opportunity to evaluate the ability of the Rosetta energy function and the DOPE Z-score to discriminate between models of varying accuracy. This is important as these are often used to pick models out of large sets of potential models. The effect of Rosetta all-atom energy minimization on model accuracy was also investigated. This was done as Rosetta all-atom energy minimization was considered as a possible step to increase model accuracy in the study set models. The energy-minimization step was considered, espe-

cially in the light of the hybrid models, as it was thought they would be likely to suffer from small, local irregularities.

4.3 Method

The aim of this chapter calls for the comparison of the accuracy of different models. To compare different models there must be an objective measure of accuracy. To accomplish such objective accuracy measurements this chapter is modelled on the CASP and Capri competitions where models are built of targets with experimentally determined structures. The models are then compared to the experimental structure using objective metrics to evaluate the similarity between the models and the experimental structure. Since the experimental structures are considered to be the correct structure, models that are more similar to them are considered better than models that are less similar.

Target complexes were identified and obtained from the PDB. Models for these complexes were built as if their structure were unknown. Many different templates were used to model each target structure. All the models were then evaluated using objective similarity metrics from the CASP and Capri projects. The results for the different templates were then compared and examined.

In an effort to get more accurate energy estimations and to improve model accuracy, all the models were subjected to intense, all atom, energy minimization using the Rosetta 3.1 relax protocol. The same evaluation metrics as before was carried out as before.

Two scoring functions, the DOPE Z-score and Rosetta energy were also evaluated by comparing the accuracy of the models to their scores according to these functions. This was done before and after the energy minimization step.

4.3.1 Targets selected

The PDB was searched for experimental structures of complexes between C1 proteases and cystatins or chagasins. Nine such structures were identified (See Section 2.4.1). These structures can be viewed in Table 2.1. There is significant redundancy amongst the targets. There are four structures of chagasin, three structures of papain, two structures of falcipain etc. Such redundancy, while problematic in the sense of reducing variation in the sample and effectively reducing the sample size, can also be useful in other respects where it allows comparison between different proteins where the binding partner stays the same. It must be kept in mind when analyzing the results.

4.3.2 Models built

The details of model building is described in Chapter 2. The creation of the templates used for modelling of the test models is described in Section 2.4.3. Self templates, simple templates

and hybrid templates were all used. The templates used for use each respective target is described in Section 2.4.4. The alignment of target and template sequences are described in Section 2.4.2. The modelling protocol is described in Section 2.4.5.

4.3.3 Calculation of scoring functions

For each model produced the DOPE Z-score and Rosetta energy was calculated as described in Section 2.4.6.

4.3.4 Evaluation metrics calculated

The models were compared to the experimental structures by calculating various evaluation metrics. All of the evaluation metrics could also be called similarity metrics, as they quantify structural similarity. Metrics were chosen from existing, published, metrics to represent both the global accuracy of the complexes as well as the accuracy of the interface region. The chosen metrics and short description of each is given below:

RMSD Standing for Root Mean Square Deviation, RMSD is the oldest and most widely used similarity metric in structural proteomics. RMSD is a type of weighted average. It is used in many disciplines. In structural proteomics it is a weighted average of the distances between corresponding positions of atoms in two different structures. RMSD upweights the larger distances. The first step is to calculate the distances for each atoms from the one position to the next. These distances are then squared, and the mean of the squares calculated. The square-root of that mean is the RMSD. The RMSD depends on the relative positioning of the two structures relative to each other. In fact, this is often used to superimpose two structures onto each other. The position that minimizes the RMSD is used as the optimal superpositioning. RMSD can be calculated for a subset of the atoms in the protein, the backbone atoms and α - carbon atoms are often used as such subsets. In this project, backbone RMSD was used. The algorithm for using RMSD to superimpose two protein structures were first described 1982 (McLachlan, 1982). Lower RMSD indicates a more accurate model. An RMSD of 0 indicates perfect similarity.

GDT-ha The Global Distance Test, roughly speaking, is an indication of what fraction of the two structures being compared is the same between the two structures. To calculate the GDT score, the fractions of atoms that fall within specific cutoff distances from the corresponding positions in the other structure is calculated. These fractions are then combined to give a single score by taking their mean. There are two main variations of GDT: GDT-ts and GDT-ha. GDT-ha uses the most stringent cutoff distances and since the models built in this project are expected to be very accurate, based on the high sequence similarity and structural conservation in the cysteine protease family, it is an appropriate choice. The cutoff distances used for GDT-ha is 0.5Å, 1.0Å, 2.0Å and

4.0Å (Zemla *et al.* , 1999). GDT scores are more robust than RMSD scores when it comes to atoms far from their correct position. In the case of RMSD, just a few atoms that are far away from their corresponding position in the other structure can lead to very bad scores. In GDT, the question is rather how many atoms are wrong, rather than how far are they off. GDT scores vary from 0 to 1 with 1 being perfectly similar and 0 being completely dissimilar.

i-RMSD Interface RMSD is the RMSD as described above, using only the backbone atoms of residues that fall in the interface of the protein complex. For the purpose of calculating i-RMSD, interface residues are defined as those residues with at least 1 atom within 10Å of an atom of the other protein in the complex. The interface residues are defined by the conformation of the target molecule, in this case, the experimental structures. The purpose of the i-RMSD score is to measure only the accuracy of the interface region, not the whole complex (Méndez *et al.* , 2003). Like with RMSD, i-RMSD must be bigger than 0 and lower values indicate higher similarity.

F-nat and F-non-nat Specific residue-residue contacts is a very important aspect of structural predictions of complexes. This information can be used for the intelligent planning of mutagenic experiments. To measure the accuracy of residue-residue contact prediction, two complimentary metrics are used. These are the fraction of native contacts (F-nat) and the fraction of non-native contacts (F-non-nat). The fraction of native contacts is the fraction of contacts in the experimental structure that are also present in the model structure. The fraction of non-native contacts is the fraction of all contacts in the model that are not present in the experimental structure. The fraction of native contacts is analogous to sensitivity and the fraction of non-native contacts is analogous to the false-positive rate in two-class predictions. For the purposes of these metrics, two residues are considered to be in contact if they have one or more atoms within 5Å of each other (Méndez *et al.* , 2003). Both F-nat and F-non-nat vary from 0 to 1. Higher F-nat values and lower F-non-nat values indicate more accurate models.

To carry out the calculations, some custom software was used in combination with publicly available software. For calculation of RMSD, the program ProFit (Martin, A.C.R. and Porter, C.T., www.bioinf.org.uk/software/profit/) was used. GDT-ha was calculated with the program TMscore (<http://zhanglab.ccmb.med.umich.edu/TM-score/>) which calculates GDT-ha in addition to several other metrics. ProFit was also used for i-RMSD, after identifying the correct interface residues using custom software written in python. F-nat and F-non-nat were calculated using custom software written in python.

4.3.5 Energy minimization carried out

To investigate the effect of Rosetta energy minimization on model accuracy, all-atom energy minimizations was carried as described in section in Section 2.4.7 on all of the models and the same evaluation metrics were calculated.

4.4 Results and discussion

An analysis of the results are given below. All models with associated scores and metrics can be obtained from the website accompanying this thesis at www.rubi.ru.ac.za/cpmdb. The results and analysis given below aim to answer specific questions and is not an exhaustive presentation of all data produced (such a presentation is not feasible).

In all the figures of this chapter, the color black is used for self models, red for simple models and blue for hybrid models. The color-key is shown in Figure 4.1 but left out of other figures as the same color scheme is used throughout. Graphs were created using the statistics program R (R-project.org).

4.4.1 General model accuracy

Before we directly compare the models to each other, it is instructive to get an idea of the general model accuracy. Table 4.1 and 4.2 shows the median RMSD of every modelgroup. It has sub-tables for each of the target complexes. Each cell represents a unique template for the target of the sub-table. The source of the protease coordinates for the template are shown on the left and the source of the inhibitor coordinates are shown at the top. The median RMSD of the modelgroup based on the self-templates are shown at the top of each sub-table. Averages are shown for each row and column with the overall average in the bottom right. RMSD was chosen as it is the most often used metric.

These tables provide a useful overview of the results. Looking at the self models we see that all of the self models have median RMSD values below 0.35Å which is roughly equivalent to X-ray diffraction structure accuracy. This shows that under ideal circumstances, the modelling protocol does not introduce a substantial amount of error.

The best result of the non-self models (simple and hybrid models) is for 1STF based on H:3E1Z-P:1NB5-I (see Section 2.4.3 for an explanation of this notation), with a median RMSD of 0.76Å. Second best, for the same target and with almost the same score of 0.77Å is the modelgroup based on the H:3IMA-P:1NB5. The accuracy can be explained by the very high similarity between the cystatin-B structure of 1STF and the cystatin-A structure of 1NB5 as well the near identical conformations of all the papain experimental structures (Figure 3.2). The templates using the same protease sources but with cystatin-A, taken from 3K9M (with cathepsin-B as protease), score slightly worse with scores of 0.85Å and 0.87Å. The cost in accuracy in taking an inhibitor from a complex with cathepsin-B instead of from a complex with cathepsin-H (which is cathepsin-L like as papain) was thus only 0.10Å. 2OUL was also modelled very well, using H:1YVB-P:3CBJ, with an RMSD of 0.77Å. This is a case of having both protease and inhibitor available in alternative experimental structures.

The least accurate modelgroup was that targeting 3K9M based on H:2OUL-P:1YVB-I and achieved a median RMSD of 11.45Å. The cause of this inaccuracy is very clearly a case of modelling cathepsin-B based on an cathepsin-L like template, resulting in the incorrect

Table 4.1 – Median RMSD values (Å) of modelgroups in the evaluation set for cystatin containing targets. The PDB ID of each target is shown in the upper left corner, inhibitor templates in the top row and protease templates on the left. Averages are shown for each row and column and the overall average shown in the bottom left corner. Median RMSD for the self-template modelgroups are shown at the top.

1NB5	Self: 0.18				Avg
	1STF	1YVB	3IMA	3K9M	
1STF	2.10	4.78	4.12	2.10	3.27
1YVB	2.12	4.52	4.04	2.06	3.18
2NQD	1.40	5.48	3.97	1.39	3.06
2OUL	2.03	5.09	4.03	2.05	3.30
3CBJ	3.70	6.60	5.30	3.65	4.81
3E1Z	2.05	5.17	4.13	2.05	3.35
3IMA	1.95	4.80	4.16	2.02	3.23
3K9M	3.81	6.27	5.31	3.83	4.80
Avg	2.40	5.34	4.38	2.39	3.63

1STF	Self: 0.28				Avg
	1NB5	1YVB	3IMA	3K9M	
1NB5	1.67	5.05	4.00	1.72	3.11
1YVB	1.97	5.02	4.12	1.99	3.28
2NQD	1.48	4.71	3.87	1.54	2.90
2OUL	2.01	5.22	4.12	2.05	3.35
3CBJ	3.30	6.14	4.92	3.28	4.41
3E1Z	0.76	4.20	3.68	0.85	2.37
3IMA	0.77	4.72	3.74	0.87	2.52
3K9M	3.57	6.23	5.04	3.55	4.60
Avg	1.94	5.16	4.19	1.98	3.32

1YVB	Self: 0.29				Avg
	1NB5	1STF	3IMA	3K9M	
1NB5	8.75	9.59	8.50	9.28	9.03
1STF	9.03	8.35	7.58	8.82	8.44
2NQD	8.69	8.40	8.40	9.67	8.79
2OUL	5.23	5.15	5.32	6.59	5.57
3CBJ	8.52	8.82	7.93	9.41	8.67
3E1Z	8.28	8.25	7.52	8.37	8.11
3IMA	9.19	8.32	8.87	9.08	8.86
3K9M	8.85	8.49	6.97	9.11	8.36
Avg	8.32	8.17	7.64	8.79	8.23

3IMA	Self: 0.18				Avg
	1NB5	1STF	1YVB	3K9M	
1NB5	2.17	2.24	2.14	2.41	2.24
1STF	1.61	1.68	1.48	1.85	1.65
1YVB	2.49	2.56	2.42	2.75	2.55
2NQD	2.05	2.13	2.05	2.34	2.14
2OUL	2.60	2.59	2.44	2.72	2.59
3CBJ	3.65	3.61	3.57	3.79	3.65
3E1Z	1.59	1.65	1.51	1.86	1.65
3K9M	3.94	3.88	3.77	4.00	3.90
Avg	2.51	2.54	2.42	2.71	2.55

3K9M	Self: 0.23				Avg
	1NB5	1STF	1YVB	3IMA	
1NB5	8.86	10.10	8.47	10.40	9.46
1STF	9.78	7.65	11.41	10.83	9.92
1YVB	10.30	8.80	10.65	11.18	10.23
2NQD	9.18	9.32	10.20	11.04	9.93
2OUL	9.29	9.51	11.45	10.78	10.26
3CBJ	2.09	2.12	4.95	3.81	3.24
3E1Z	9.39	10.11	10.55	7.99	9.51
3IMA	8.52	7.71	10.82	10.41	9.37
Avg	8.43	8.16	9.81	9.55	8.99

Table 4.2 – Median RMSD values (Å) of modelgroups in the evaluation set for chagasin containing targets. The PDB ID of each target is shown in the upper left corner, inhibitor templates in the top row and protease templates on the left. Averages are shown for each row and column and the overall average shown in the bottom left corner. Median RMSD for the self-template modelgroups are shown at the top.

2NQD	Self: 0.172			
	2OUL	3CBJ	3E1Z	Avg
1NB5	1.51	1.44	1.73	1.56
1STF	1.90	1.84	2.03	1.92
1YVB	2.07	1.97	2.25	2.10
2OUL	2.15	2.02	2.26	2.14
3CBJ	3.25	3.21	3.28	3.25
3E1Z	1.89	1.87	1.95	1.91
3IMA	1.85	1.83	2.17	1.95
3K9M	3.45	3.41	3.55	3.47
Avg	2.26	2.20	2.40	2.29

2OUL	Self: 0.34			
	2NQD	3CBJ	3E1Z	Avg
1NB5	6.09	5.67	6.21	5.99
1STF	5.89	6.67	5.64	6.07
1YVB	1.05	0.77	0.83	0.88
2NQD	5.98	6.01	5.47	5.82
3CBJ	5.99	6.22	5.10	6.07
3E1Z	6.46	5.96	7.02	6.48
3IMA	6.84	6.60	6.61	6.68
3K9M	7.05	6.40	7.27	6.91
Avg	5.669	5.54	5.63	5.61

3CBJ	Self: 0.20			
	2NQD	2OUL	3E1Z	Avg
1NB5	8.35	8.26	9.21	8.61
1STF	8.56	8.12	9.45	8.71
1YVB	8.97	9.06	9.08	9.04
2NQD	7.87	8.15	8.19	8.07
2OUL	9.04	8.99	9.10	9.04
3E1Z	7.58	7.57	9.75	8.30
3IMA	9.27	7.37	9.46	8.70
3K9M	2.29	2.22	2.29	2.27
Avg	7.74	7.47	8.32	7.84

3E1Z	Self: 0.23			
	2NQD	2OUL	3CBJ	Avg
1NB5	1.90	1.71	1.83	1.81
1STF	1.32	0.88	1.08	1.09
1YVB	2.17	2.02	2.07	2.09
2NQD	1.76	1.52	1.64	1.64
2OUL	2.19	2.00	2.07	2.09
3CBJ	3.47	3.25	3.28	3.33
3IMA	1.44	0.88	1.04	1.12
3K9M	3.75	3.57	3.53	3.61
Avg	2.25	1.98	2.07	2.01

modelling of the occluding loop; in addition, the modelling of a type I (cystatin-A) based on a type II cystatin (ovocystatin) leads to the template-free modelling of the C-terminal extension.

Many interesting deductions can be made purely by looking at the values in the tables. For example, by looking at the results for models of 1YVB (falcipain-2 and ovocystatin) it may not be immediately clear if the protease or the inhibitor are the main sources of error. However, by comparing the results to models targeting 2OUL we can conclude that the error is mainly present in the inhibitor. This is because the excellent models of 2OUL based on the 1YVB protease (RMSD as low as 0.77Å) shows that the 1YVB and 2OUL proteases are very much alike, yet even the best modelgroups of 1YVB, based on the 2OUL protease, have a relatively high median RMSD, at 5.32Å. The source of the error therefore is clearly the inhibitor and most likely results from the template-free modelling of the insert on side of the molecule opposite the interaction site. This region is absent in all the other cystatin structures, including tarocystatin, the only other type II cystatin in the template set (Figure 3.7).

To further summarize the general model accuracy, the distributions of the RMSD, GDT-ha, i-RMSD, F-nat and F-non-nat, for all models built in the test set, is shown in Figure 4.1. Since F-nat and F-non-nat are closely related metrics, they are shown together on a scatterplot instead of individually on a histogram like the other metrics.

The models cover a large range of accuracy. The RMSD and i-RMSD scores go up to 12Å. The i-RMSD scores are mostly clustered around 1.5Å with a long tail to the right. The RMSD scores show a more flat distribution with a peak around 2.5Å. The GDT-ha scores vary from 0.35 to 1. F-nat varies from 0.4 to 1 and F-non-nat from 0 to 0.5.

The i-RMSD results show that the interface region of the proteins are generally more reliably modelled than the global protein structures. The main contributor to large i-RMSD values is the template free modelling of the cathepsin-B occluding loop when using cathepsin-L like templates. Other factors include the modelling of cathepsin-L like models using cathepsin-B and the modelling of type I cystatins using type II cystatin templates. If these 3 factors are avoided, then all models have i-RMSD below 2.5Å. Global RMSD is affected by the same factors, but additionally by the template free modelling of the nose and arm regions in falcipain-2 and the template free modelling of the loop region in ovocystatin on the non-interface side of the molecule.

The GDT-ha results give an indication of the side-chain accuracy and the self models are clearly separated from the other models. Some hybrid models were built using identical proteins to the target complex for both the protease and inhibitor. Such models give RMSD results close to self models but their GDT-ha scores are clearly separated from the self models. This indicates that errors are mainly associated with the side-chains.

In the case of F-nat and F-non-nat there is also a very clear separation of self models and other models. As with GDT-ha this is likely due to side-chain sensitivity.

In all the metrics, the hybrid and simple models show similar distributions.

It is informative to get an idea of the variation in model quality within the modelgroups. To summarize this, Figure 4.2 shows a scatterplot of the standard deviation of the RMSD, of

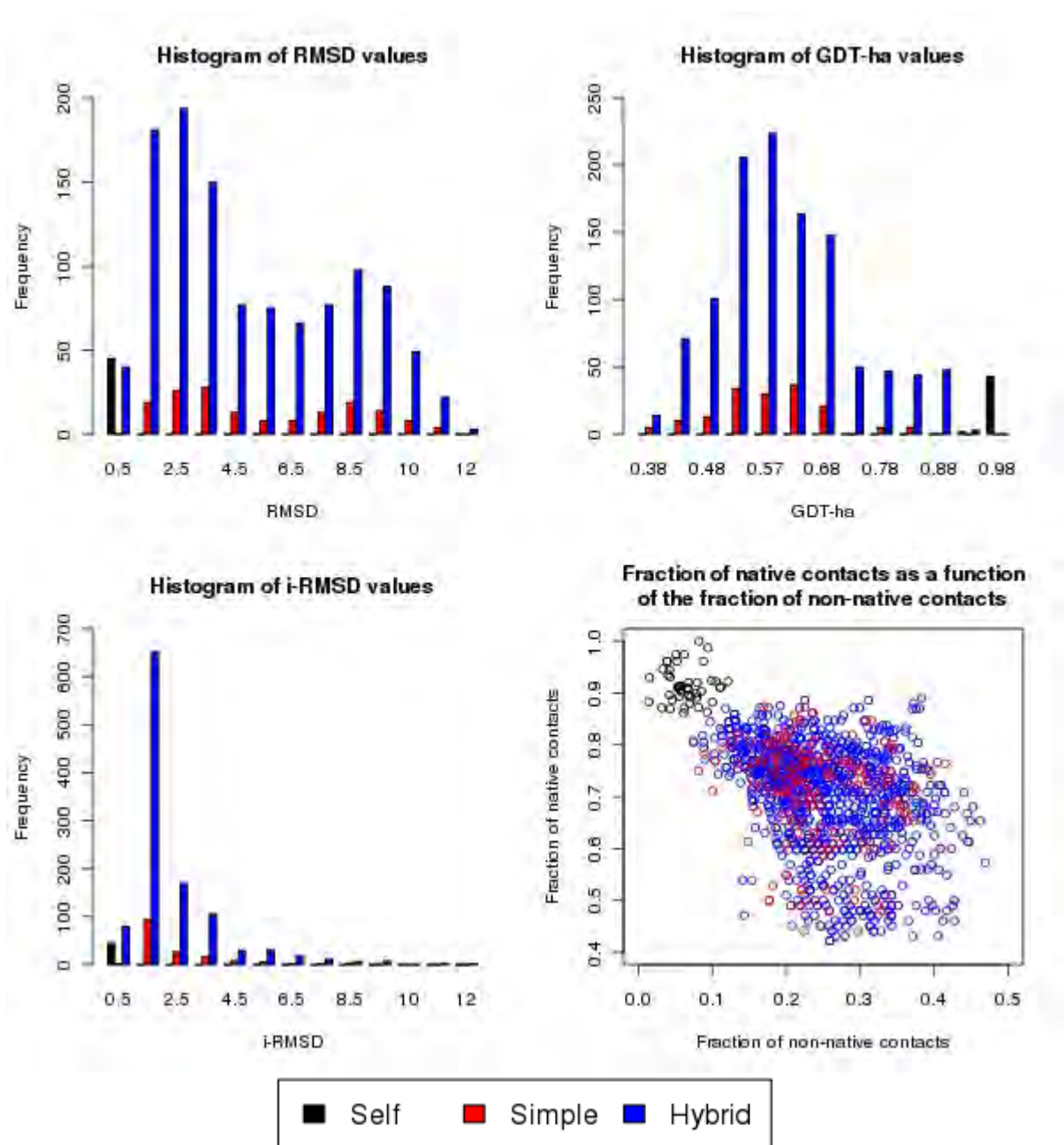


Figure 4.1 – Distribution of major evaluation metrics for all test set models.

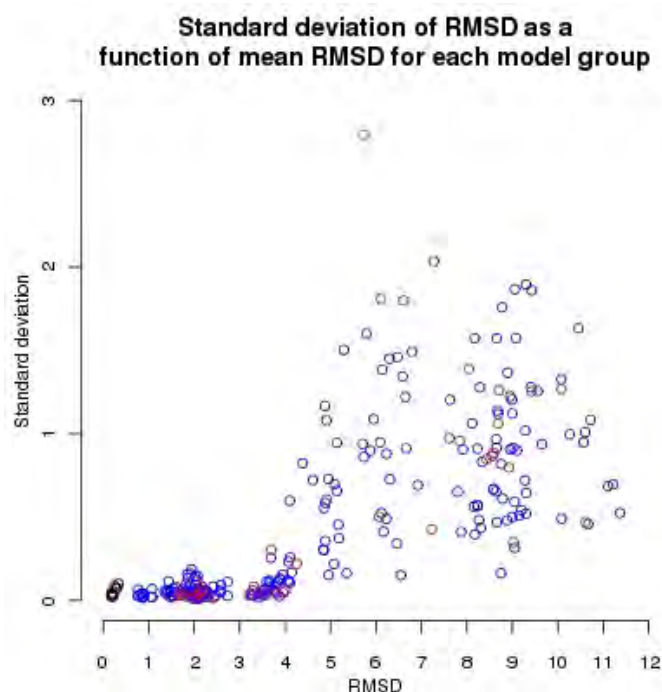


Figure 4.2 – Standard deviation of RMSD values within a modelgroup as a function of the mean RMSD of that modelgroup.

the models within each modelgroup, as a function of the mean RMSD of that modelgroup. The result is very interesting. There is a clear and sharp increase in model variation starting from about 4Å upwards. This is a comment on the method used to produce the models and suggests that the program correctly recognizes accurate conformations and does not change them significantly. In the presence of uncertainty or low confidence, such as the presence of a loop, the program introduces more variation in an attempt to find a more accurate conformation. The variation between the models produced is thus an indication of the accuracy of the models. However, this correlation is by proxy and there should be better ways to estimate the accuracy of the models through examination of sequence similarity. In addition, the variation of the models only places them in one of two categories of having RMSD greater or smaller than 4Å. It is still very interesting to see the shape of the graph. It is not clear why there is such a sharp increase instead of a more gradual rise.

4.4.2 Direct comparison of simple and hybrid templates

The act of creating hybrid templates (as described in Section 2.4.3) could potentially introduce irregularities into the templates and these could carry over into the models. To test whether this was the case, a subset of the test set, in which the model and hybrid templates had the same distribution of sequences similarity to the targets, was analyzed. The major similarity metrics were directly compared to each other using the Mann-Whitney U-test (Mann

		1YVB	1STF	3IMA	1NB5	3K9M	2OUL	2NQD	3CBJ	3E1Z
		Ovocyst	Cyst-B	Tarocyst	Cyst-A	Cyst-A	Chag	Chag	Chag	Chag
1YVB	FP-2	S	H	H	H	H	H	H	H	H
1STF	Papain	H	S	H	H	H	H	H	H	H
3IMA	Papain	H	H	S	H	H	H	H	H	H
1NB5	Cath-H	H	H	H	S	H	H	H	H	H
3K9M	Cath-B	H	H	H	H	S	H	H	H	H
2OUL	FP-2	H	H	H	H	H	S	H	H	H
2NQD	Cath-L	H	H	H	H	H	H	S	H	H
3CBJ	Cath-B	H	H	H	H	H	H	H	S	H
3E1Z	Papain	H	H	H	H	H	H	H	H	S

Figure 4.3 – Subset of test set used for the direct comparison between simple models and hybrid models, only structures containing cystatin were considered.

& Whitney, 1947).

Only complexes with cystatin inhibitors and only proteases that are from crystal structures with cystatin were included in the subset (Figure 4.3). This insures identical sequence similarity distributions across both classes of models. While this causes a big decrease in total number of models, only 2 test set proteins are completely excluded from the set. These are chagasin and cathepsin-L. To see why this selection ensures identical similarity distributions, refer to Figure 4.3. The exclusion of chagasin limits the target complexes to those containing cystatin. That leaves five targets, each modelled using four simple templates and twelve hybrid templates obtained from the other four targets. Thus, for modelling the target protease, four template proteases are available. Each of the template proteases is used once in a simple template and three times in a hybrid template. This fixed ratio of 1:3 ensures that the distributions are the same. For a single target, every time a specific protease is used in a simple template it is used three times in a hybrid template. A similar argument can be made for the inhibitors. Thus, for each target, the distribution of sequences used is exactly the same between the the simple templates and hybrids. Since all targets have the same number of templates, this equality is preserved when pooling them all together to give the comparison subset.

The Mann-Whitney U-test allows the comparison of unequal samples with non-normal distributions. It uses a statistic based on the ordering of results from two sets to calculate the probability of encountering the observed difference between two sets as a result of random sampling error. This probability is called the P-value. The P-value also depends on what alternative hypothesis is used. The alternative hypothesis is the alternative to the null hypothesis. The null hypothesis is that the two samples are the same, in terms of the measured quality. The most common alternative hypothesis is simply that the two samples differ. A one-sided alternative hypothesis assumes that any difference between the two sets can only

be in one direction. In other words it assumes that the measurements of set A may be larger than the those of set B, but the measurements of set B cannot be larger than those of set A. The alternative hypothesis used, in the present comparison, was that the simple models are more accurate than the hybrid models. Thus, a one-sided alternative hypothesis was used. This was done as there is no reason to suspect that a hybrid model would be better than a simple model, given the same sequence relationships between the template and target proteins. Figure 4.4 shows the distributions of the major evaluation metrics, for the representative set only, along with the differences between the means of the sets and the P-values resultant from the Mann-Whitney tests.

The results support the null-hypothesis - that there is no significant difference between the simple and hybrid models in terms of the major evaluation metrics. The difference between the means of the metrics are very small in any case. The metric with the most significant difference is F-non-nat with a P-value of 0.13 and a difference of 0.0095 between the means. It is interesting that the i-RMSD and RMSD scores are actually lower (by an insignificant amount) in the hybrid models than in the simple models. The GDT-ha score for the hybrids is smaller by an insignificant amount.(P-value of 0.31)

The result is quite interesting as it shows that hybrid models are of the same quality as simple models. Irregularities in the structure, produced in the hybrid construction process, are either insignificantly small or are overcome by the modelling protocol. On the practical side, the result means that there is no loss in accuracy when using hybrid templates. When selecting templates for the modelling of further complexes, no compromise between the inhibitor and protease is necessary. The best template for the protease and the best template for the inhibitor can both be used without introducing errors.

4.4.3 Examination of relation between model accuracy and sequence identity and coverage

This section investigates the relationship of sequence identity and coverage with model accuracy.

Sequence identity is the percentage of amino acids between two sequences that are conserved. Positions in the sequence, where either sequence has a gap in the alignment, are ignored. Sequence identity is a measure of the evolutionary distance between two proteins. The higher the identity, the more recent is the common ancestor of the sequences. There is a correlation between sequence identity and structural similarity (Chothia & Lesk, 1986). Since structural similarity between the target and template proteins is the main factor influencing the accuracy of a homology model, sequence identity is a good indication of the accuracy that can be expected of a model.

Sequence coverage is the percentage of amino acids in a sequence which are not aligned to gaps. The sequence coverage is thus a measure of the amount of "gaps" in the other sequence. In the context of homology modelling, it can be used as a measurement of how much template

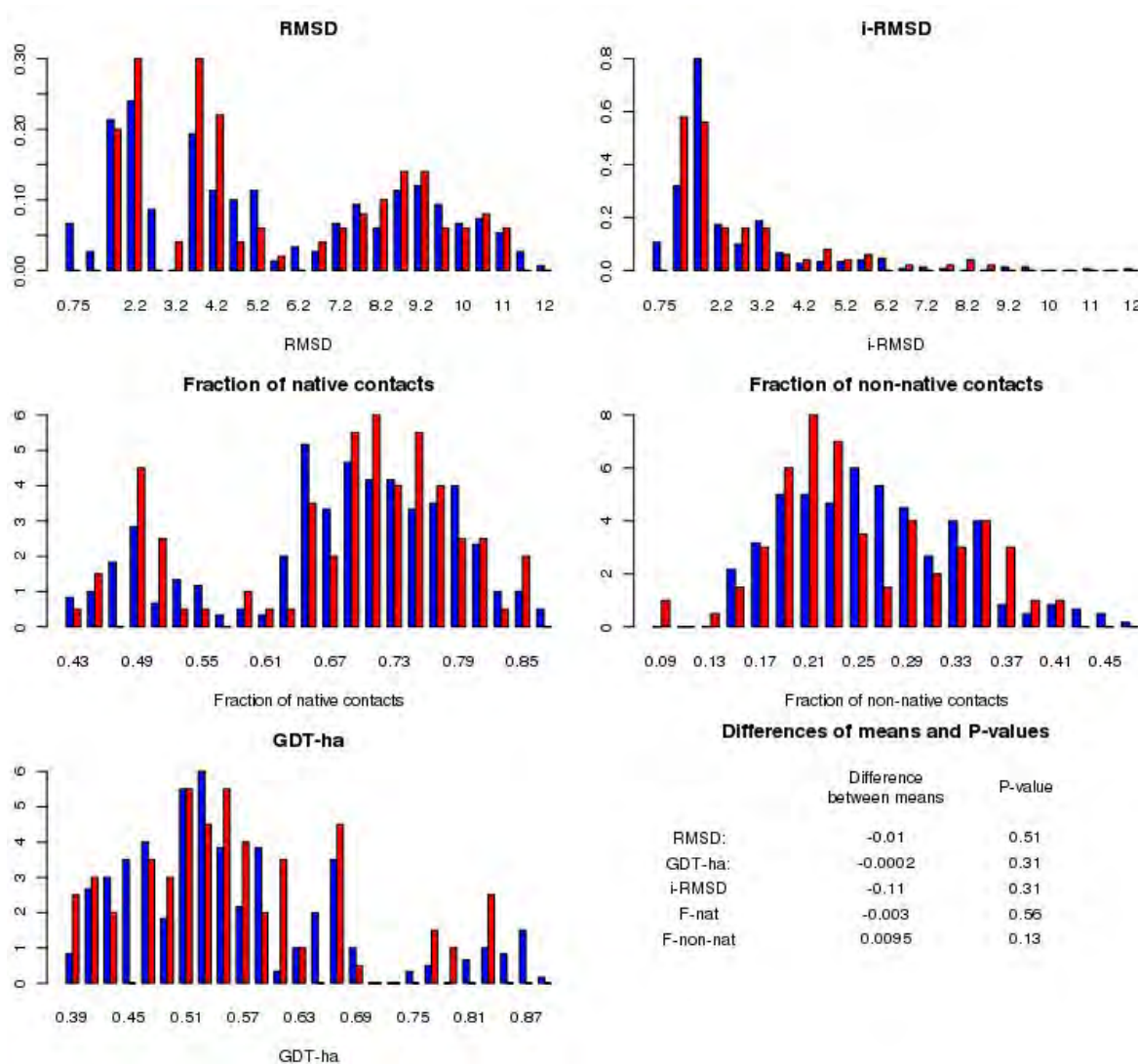


Figure 4.4 – Comparison of major evaluation metrics between simple and hybrid models of a subset of the test set. Distributions, differences between means, and Mann-Whitney U-test P-values of the major evaluation metrics are shown. The difference in means were calculated by subtracting the means of the simple models from the means of the hybrid models. The graphs show histograms with a normalized frequency on the y-axis and the values of the evaluation metric itself on the x-axis.

free modelling has to be done for a target protein. One can say that a model was built at 80% coverage, meaning that 80% of amino acids of the target protein was aligned to a template structure. Since template free modelling is a major source of errors in structural models, the sequence coverage is an important measurement that is correlated with the accuracy of homology models.

To elucidate the relationship between sequence identity and coverage with model accuracy, RMSD and i-RMSD were investigated. RMSD serves as a measure of global accuracy while i-RMSD is a measure of the interface accuracy.

Figure 4.5 shows the median RMSD of the test set models as functions of inhibitor identity, protease identity, inhibitor coverage and protease coverage respectively. Each target complex is shown as a separate data series. Where inhibitor identity or coverage is shown on the x-axis, the RMSD was averaged over the various proteases used. Similarly, the average over the various inhibitors was taken where protease identity or coverage is shown on the x-axis. In this way, the influence of inhibitor and protease error on the RMSD can be separated. The slope of the lines carry the information in these graphs. The absolute value of the series is unimportant as it is influenced by the contribution of the other complex component, not the one being investigated.

It is interesting to see that increasing inhibitor sequence identity from 50% to 100% does not seem to improve the model RMSD. This shows that the structure of the inhibitor (in this case cystatin) is still highly conserved at the level of 50% sequence identity. In the case of protease sequence identity, it can be seen that the relation is not consistent and depends on the specific case under consideration. The trend is always that higher sequence identity leads to lower RMSD values, however, by how much it is lowered, depends on the specific case. There is some unexpected results when looking at the sequence coverage of the inhibitors. There is clearly an increase in RMSD (decrease in accuracy) associated with an increase in coverage in the section of the graph between 80% and 90% coverage. The targets visible there are all type I cystatins and the two inhibitor templates are 1YVB - ovocystatin and 3IMA - tarocystatin. The coverage provided by ovocystatin is higher, although its structure is farther from that of the type I cystatins. The protease show a much clearer and more consistent downward trend with regards to sequence coverage.

Figure 4.6 was constructed in exactly the same way as figure 4.5 except that it shows i-RMSD instead of normal RMSD. The results, again, are interesting. The correlation between i-RMSD and inhibitor coverage is very strong. It is also surprising that some of the best models in terms of i-RMSD was built at the lowest levels of coverage. This is explained by the fact that the gaps do not occur in the interacting region. For the proteases it seems that the strongest correlation is with sequence identity.

Overall, the results support the theoretical position that higher sequence identity and higher sequence coverage is correlated with more accurate models. However, the results also show there is no one-to-one relationship between these metrics and accuracy. Depending on the specific case, identity or coverage may be more important and the impact may vary. Structural knowledge of the families serving as templates and targets is useful in interpreting the results

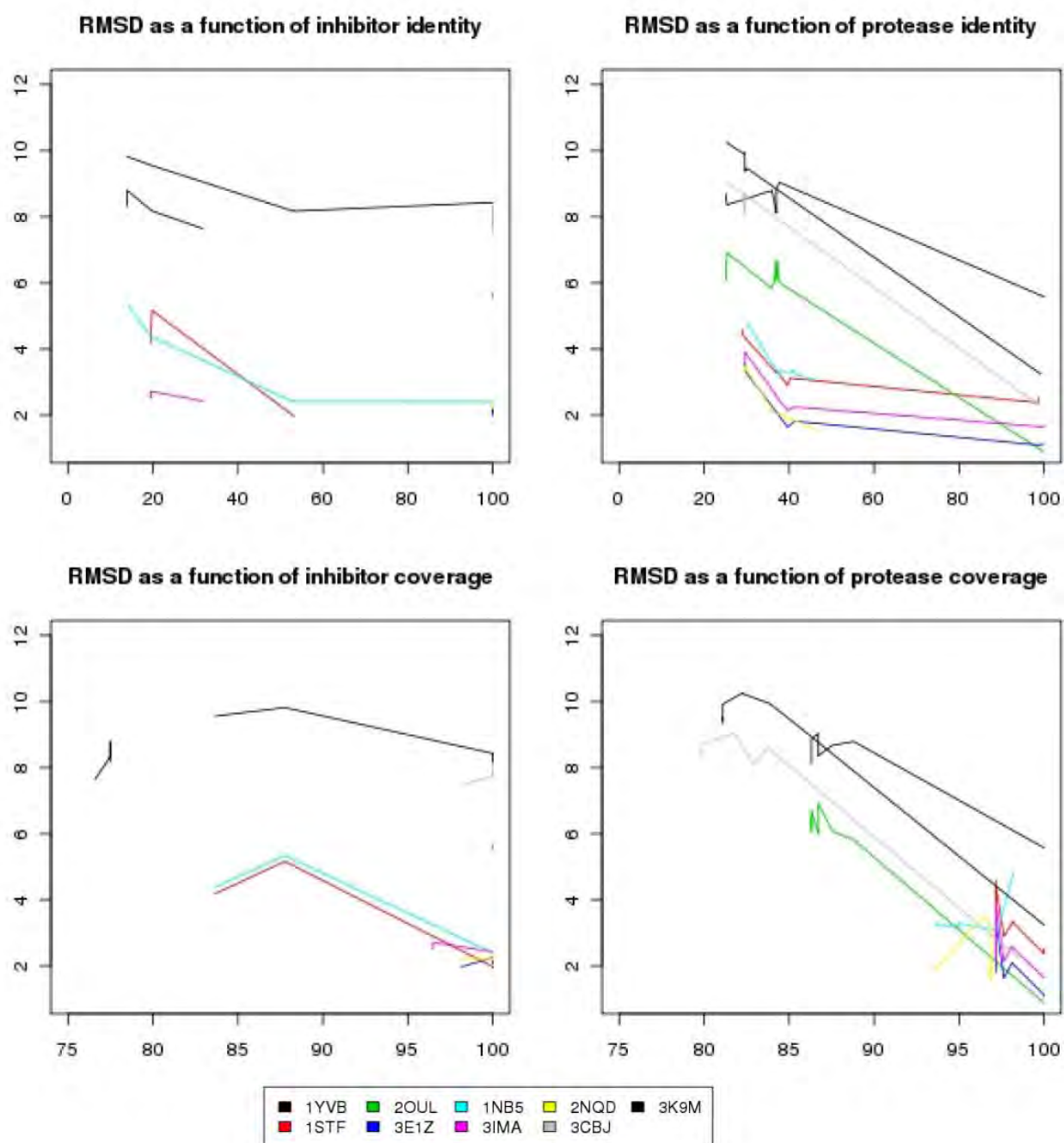


Figure 4.5 – Plots showing the relationship between RMSD, sequence identity and sequence coverage for each of the 9 targets.

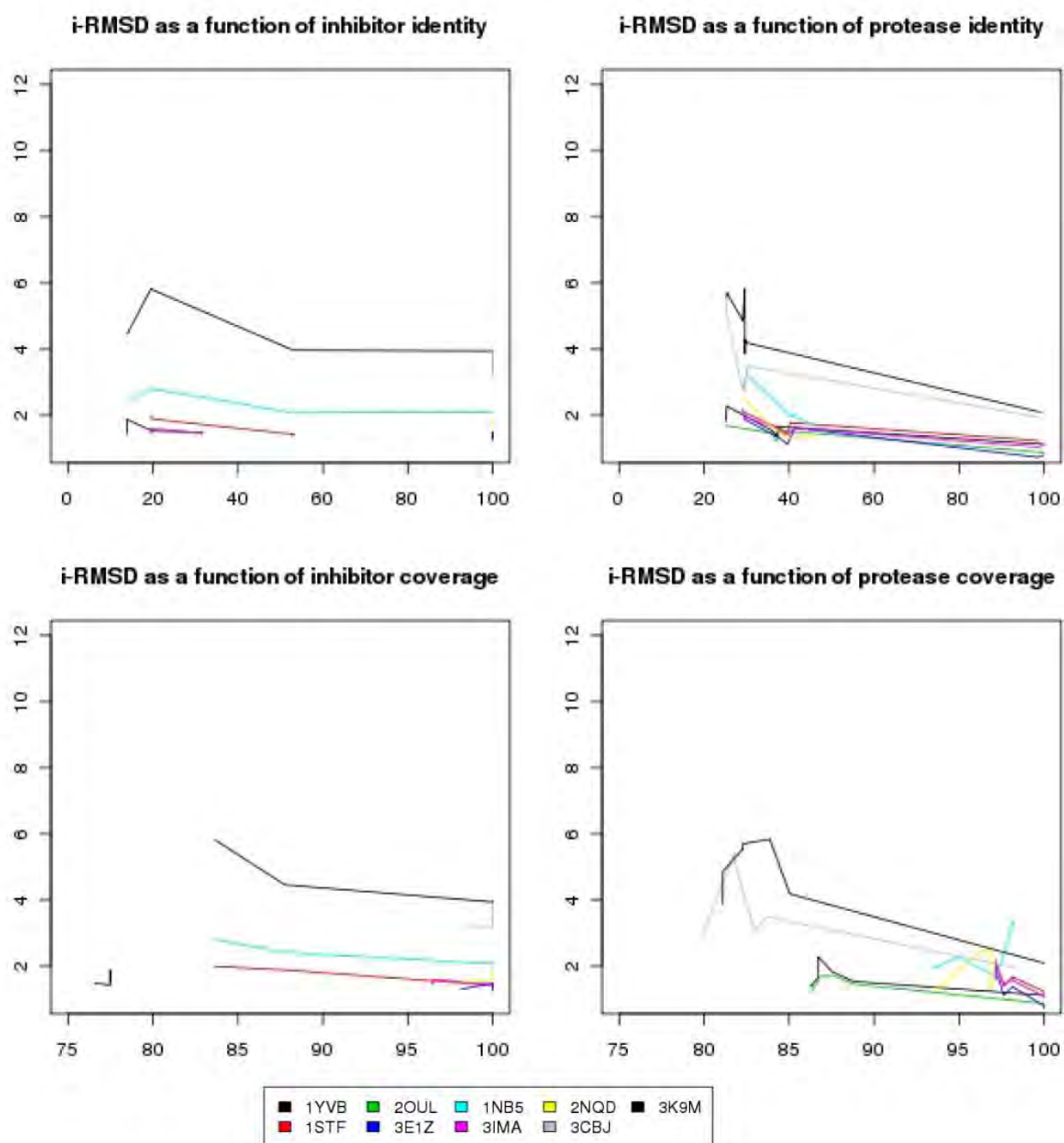


Figure 4.6 – Plots showing the relationship between i-RMSD, sequence identity and sequence coverage for each of the 9 targets.

and selecting templates.

4.4.4 Examination of relationship between scoring functions and model accuracy

An often used method for increasing model accuracy is to build many models and then select the best one to use further. A scoring function, in the protein modelling context, can be used to refer to any mathematical function that has as input a protein structure and outputs a value that aims to be an indication of the accuracy of the model. Physical energy functions and statistical potentials are often employed as scoring functions, to select the best models from a set of potential models. Two scoring functions, the Rosetta energy function and the DOPE Z-score were tested for their ability to identify accurate models. The scoring function values were calculated for all the test models and compared to the actual accuracy of the models, as measured using RMSD, to evaluate their performance.

Figure 4.7 shows a positive correlation between the RMSD and Rosetta energy. A lower Rosetta energy typically corresponds to a lower RMSD value and in all cases the best (lowest) scoring model is one of the most accurate models. The correlation is not very strong. For a given energy score, the corresponding RMSD value can vary quite substantially. Some results seem to defy the correlation completely. Specifically, some models of complex 1STF have very low energy scores despite having double the RMSD of other models with much higher energy values. In conclusion, the Rosetta energy does show a clear, positive, correlation with RMSD and a low energy model is thus likely to be more accurate than a high energy model.

Figure 4.8 shows a similar positive correlation between the DOPE Z-score and RMSD values. Inspection indicates fewer anomalous scores than with Rosetta energy. The models for 1STF still contains a subset of models that are scored too low (with respect to RMSD) but they are, correctly, scored significantly worse than the best cluster, unlike with the Rosetta energy. In other cases like 1NB5 and 3K9M, DOPE Z-score also show better performance than Rosetta energy as the best models are scored lower than the others.

Both scoring functions clearly show the ability to make better than random selections out of a set of models built on different templates, with DOPE Z-score showing slightly better performance. The next question is if they are able to identify the best model within a modelgroup. This is very important. The reason for building multiple models, using the same template, is that the chance of building an accurate model is increased through multiple attempts. If the most accurate model cannot be identified, it creates a situation where the researcher cannot determine which model to use for further work and the benefit of building multiple models is largely lost. In the best case scenario, the researcher can carry on research with the entire ensemble of models, but this is currently not practical in most cases.

To evaluate the ability of the two scoring functions to identify the most accurate models in a modelgroup, the score of each model, relative to the mean score of its modelgroup, was plotted as a function of its RMSD, relative to the mean RMSD of its modelgroup. In other

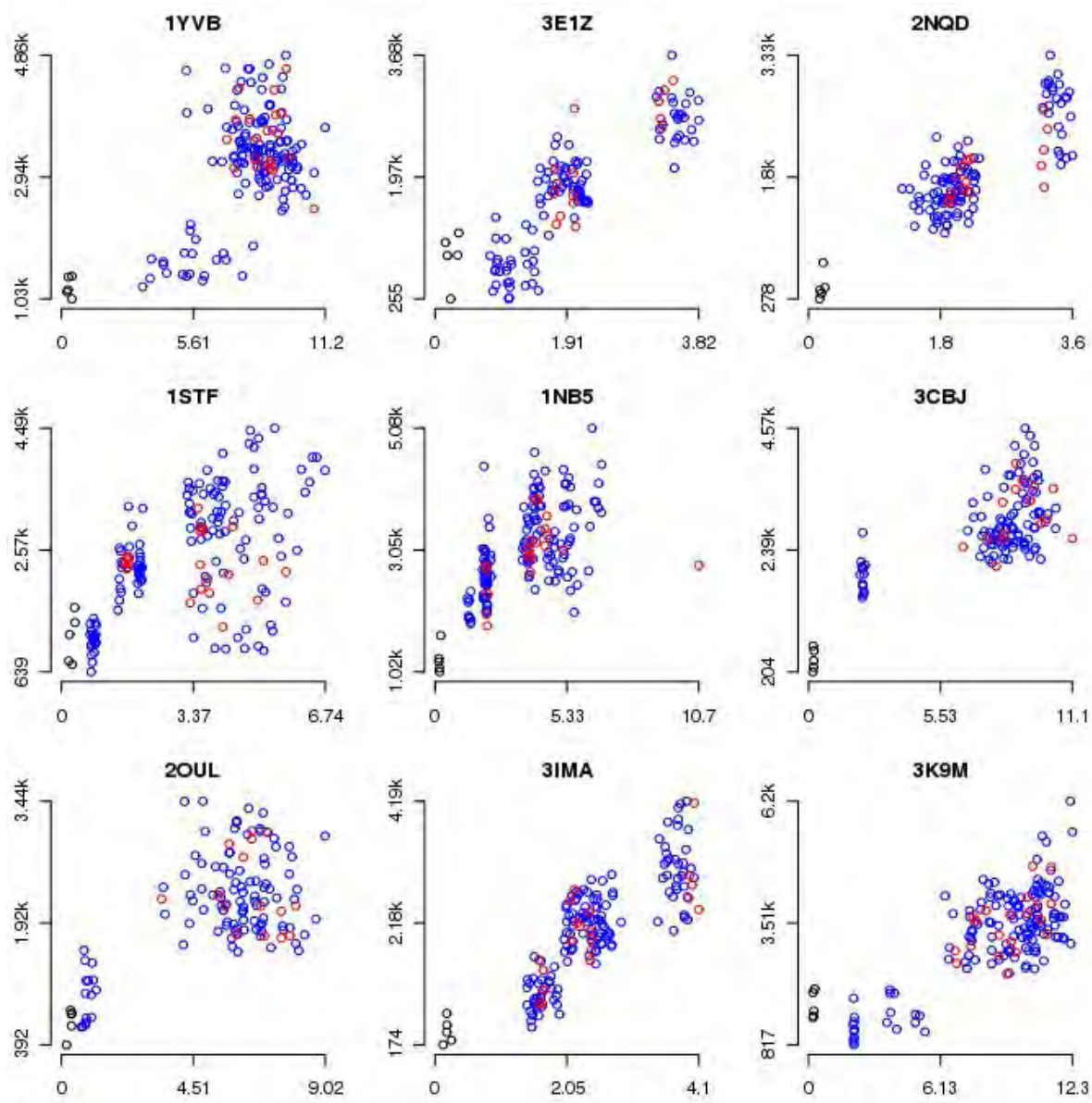


Figure 4.7 – Rosetta energy as a function of RMSD for all models in the test set. Each target complex is shown in its own sub-graph.

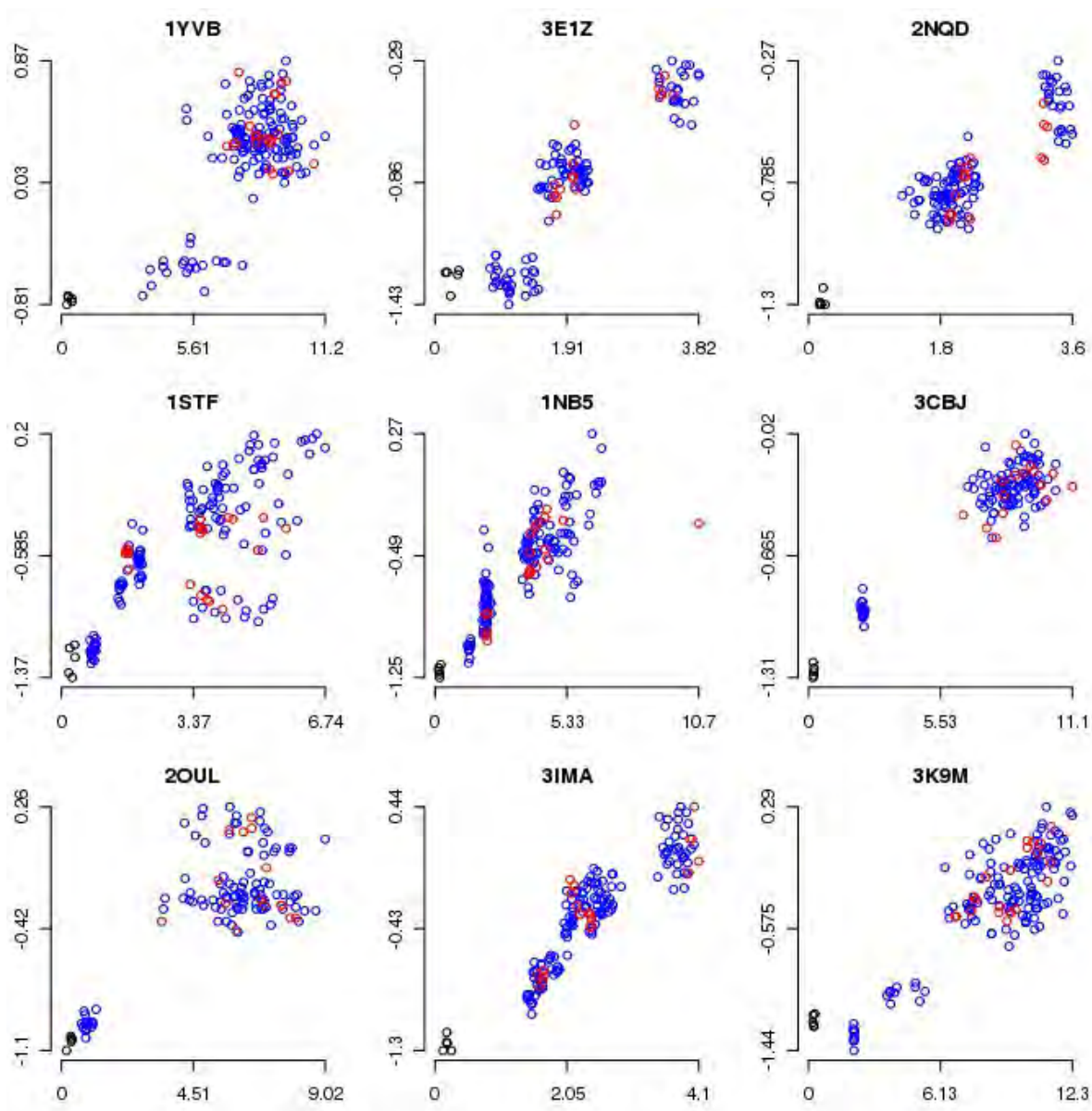


Figure 4.8 – DOPE Z-score as a function of RMSD for all models in the test set. Each target complex is shown in its own sub-graph.

words, the shift in score relative to the modelgroup was plotted as a function of the shift in RMSD relative to the modelgroup. If the scoring functions are successful, accurate models (which have a lower RMSD than the mean RMSD of the modelgroup) should have scores lower than the mean score for that modelgroup (in both Rosetta energy and DOPE Z-score, lower scores correspond to more accurate models) and thus would appear on the bottom left of the graph. Similarly, inaccurate models should appear on the top right of the graph. A successful scoring function would thus result in positive correlation with the results forming a trend from the bottom left to the top right.

In Figure 4.9, with the exception of the self models, no such trend is visible for either Rosetta energy or DOPE Z-score. The Pearson Correlation coefficients also indicate that the scoring functions fail with the simple and hybrid models. In contrast, it seems that both scoring functions are actually able to detect the more accurate models amongst the self models. From this result, it seems possible that the scoring functions are in fact able to identify the most accurate models amongst already very accurate models and that this ability is hidden by the less accurate models amongst the simple and hybrid models.

To examine whether the performance of the scoring functions varies according to the accuracy of the model, the Pearson correlation coefficient between the scoring functions and RMSD within each modelgroup was plotted as a function of the mean RMSD of that modelgroup. In this case a positive correlation indicates good performance. The lines indicate the mean value of the datapoints in 1Å bins. In other words, the line indicates the mean Pearson Correlation coefficient achieved within a certain range of accuracy. The accuracy ranges used were 0.0Å - 1.0Å, 1.0Å - 2.0Å, 2.0Å - 3.0Å. etc.

Figure 4.10 supports the conclusion that the scoring functions are unable to distinguish between models within a modelgroup. For very accurate models (below 2Å) the results is inconclusive but even if there is a slight correlation between accurate models and good scores, it is very weak. Overall, the deviations from zero show no pattern and are thus likely due to random chance. The biggest deviations from zero occur where there are few data points, making them unreliable. It is also better to look at the hybrid models as there are more data points and thus higher confidence. In the case of the self models, the scoring functions definitely show the ability to correctly identify accurate models (better than random). A possible explanation for this is that the scoring functions take the side-chains into account, while they are not reflected in the RMSD scores. In the self models, the side-chains are more likely to be correct and thus less likely to skew results of the scoring functions.

4.4.5 Examination of minimization results

4.4.5.1 Effect on accuracy

To examine the effect of the energy minimization on accuracy, the changes in the major evaluation metrics are shown in Figure 4.11 as a function of their value before the minimization.

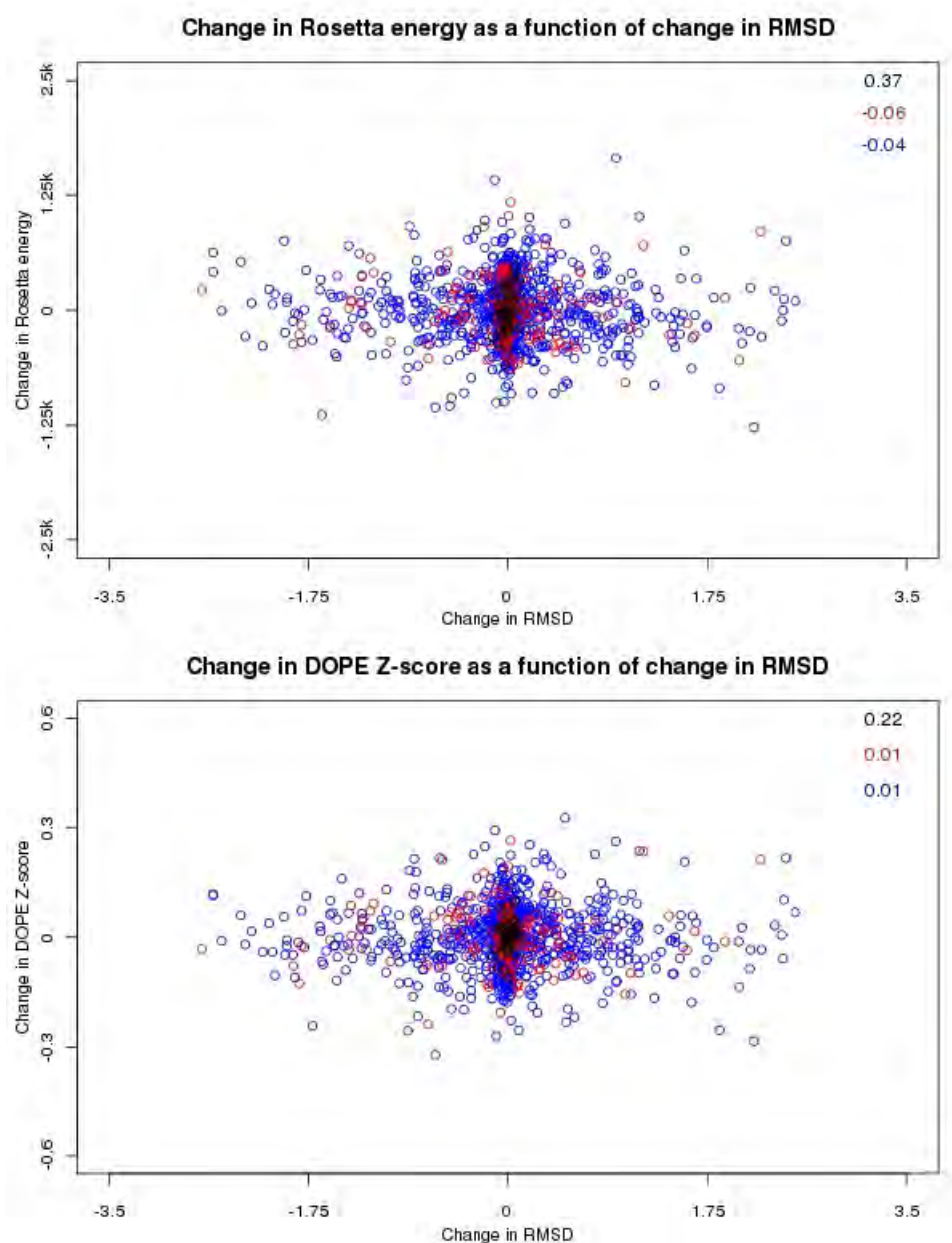


Figure 4.9 – The score of each model, minus the mean score of the modelgroup, is plotted against the RMSD value of that model, minus the mean RMSD of the modelgroup. The numbers in the corner of the graphs indicate the Pearson correlation coefficient between the scoring function and RMSD in each case for self models, simple models and hybrid models.

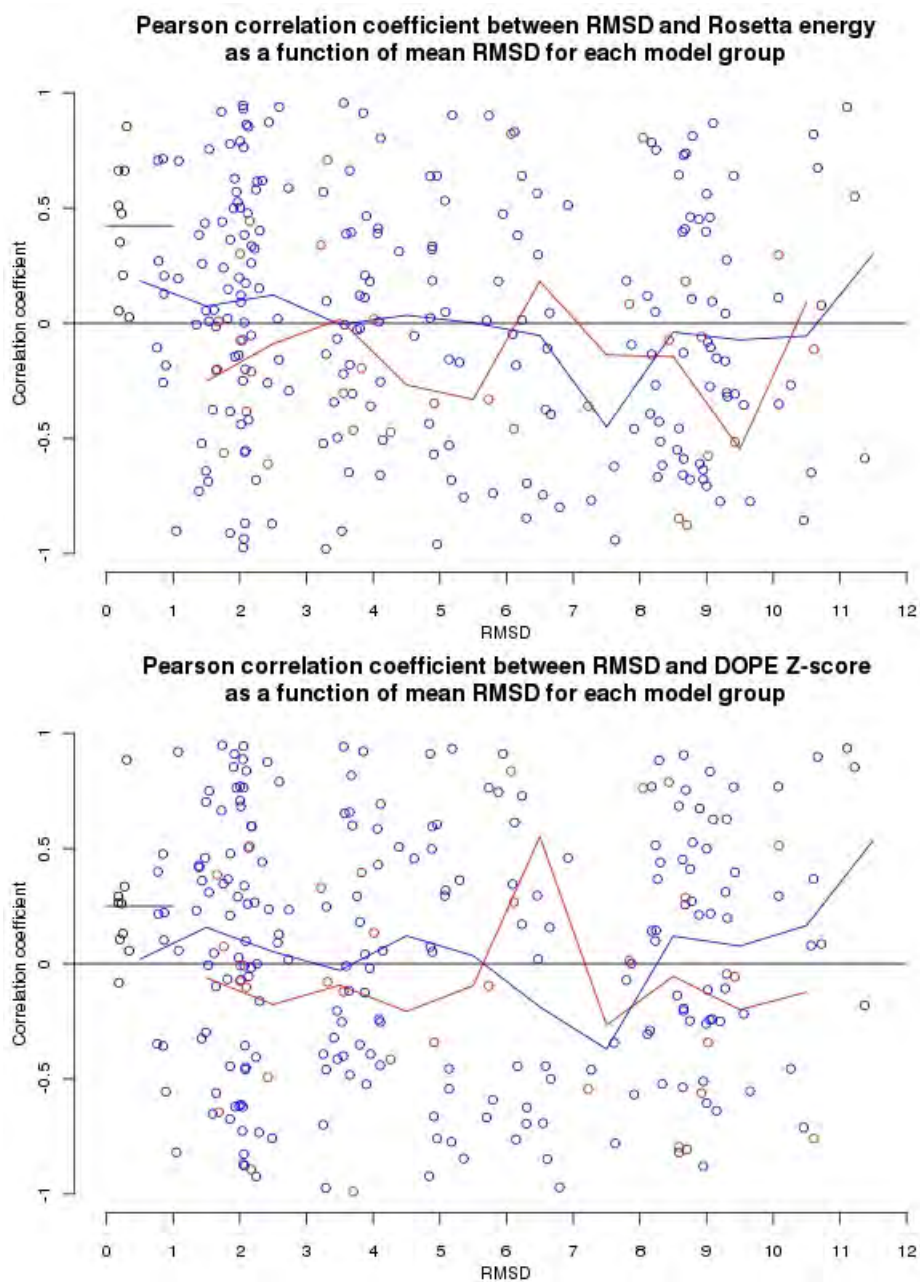


Figure 4.10 – The circles indicate the Pearson correlation coefficient between the indicated scoring functions and RMSD within each modelgroup is shown as a function of the mean RMSD of the modelgroup. The lines indicate the average value of each series over intervals of 1 Å.

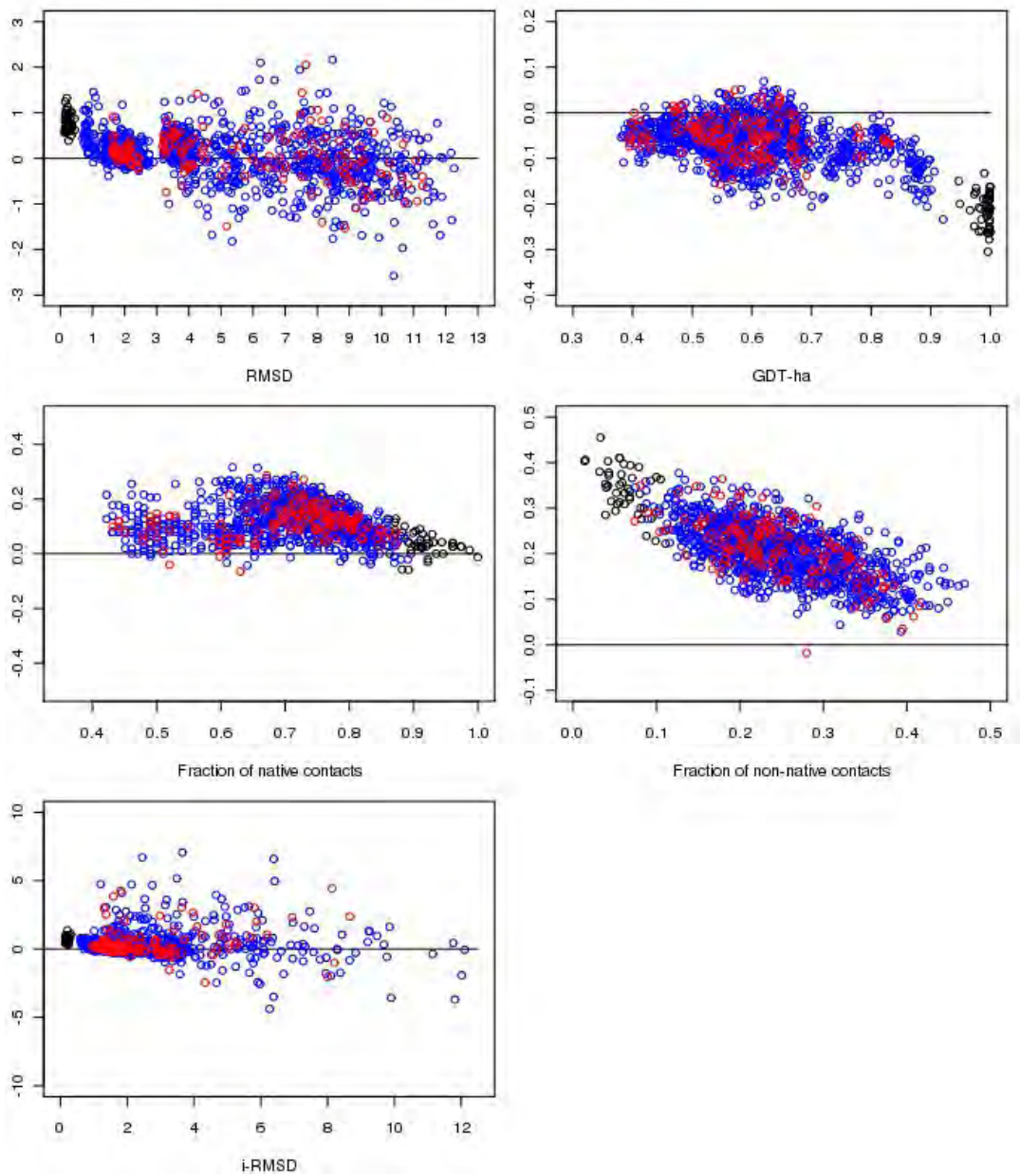


Figure 4.11 – Changes in evaluation metrics due to minimization is shown as a function of the value before minimization.

This was done as there was reasonable suspicion that the effect might be dependent on the accuracy before minimization.

From the figure it is clear that minimization leads to a decrease in accuracy (in terms of the evaluation functions) in models that are already very accurate. This is expected as any change in a highly accurate model is likely to reduce the accuracy. It is noteworthy that the loss in measured accuracy is only slightly larger than the variation between different crystal structures of the same proteins. The minimization is thus not necessarily making the model more different from the native protein structure, but different from the crystallized conformation.

For the less accurate models, it seems that roughly the same number are improved as degraded. This is very impressive and could be very useful if improved models could be reliably detected afterwards (without having the experimental structure available as it is here) so that improvements can be retained while degradations are discarded. Unfortunately, according to findings of section 4.4.4, that is not possible.

GDT-ha seems to be reduced in the majority of cases. This is likely due to the already accurate models and the sensitivity of GDT-ha to side-chain conformations. The fraction of native contacts was improved almost universally while the fraction of non-native contacts was worsened. This implies that the total number of contacts predicted for each model was increased and the protease and inhibitor were brought into closer contact than before. The i-RMSD scores worsened in the case of already low scores but there seems to be a trend toward improvement for less accurate models.

4.4.5.2 Are the energy scores more accurate for the minimized models?

Since the scoring functions, especially the Rosetta energy, can be sensitive to small, local, irregularities, it is possible that their performance would be better with minimized models in which such irregularities should theoretically be removed. As before, the Pearson correlation coefficient between the scoring functions and the RMSD values after minimization within each modelgroup is shown as a function of the mean RMSD of that modelgroup. The lines indicate the mean of the datapoints in each series in 1Å bins.

Figure 4.12 shows a very small, but real, improvement over Figure 4.10 for the DOPE Z-score. The DOPE Z-score now clearly has a small but positive correlation as the mean correlation coefficient is above 0 far more than below. This correlation is still very small and selecting the top-scoring models will only provide a slight enrichment of accurate models in the selection.

4.4.5.3 Examination of relation between the energy change during minimization with the accuracy of the model

Since a good model should already have a low energy it should not, on average, improve as much as a less accurate model. The change in the scoring functions are shown as a function

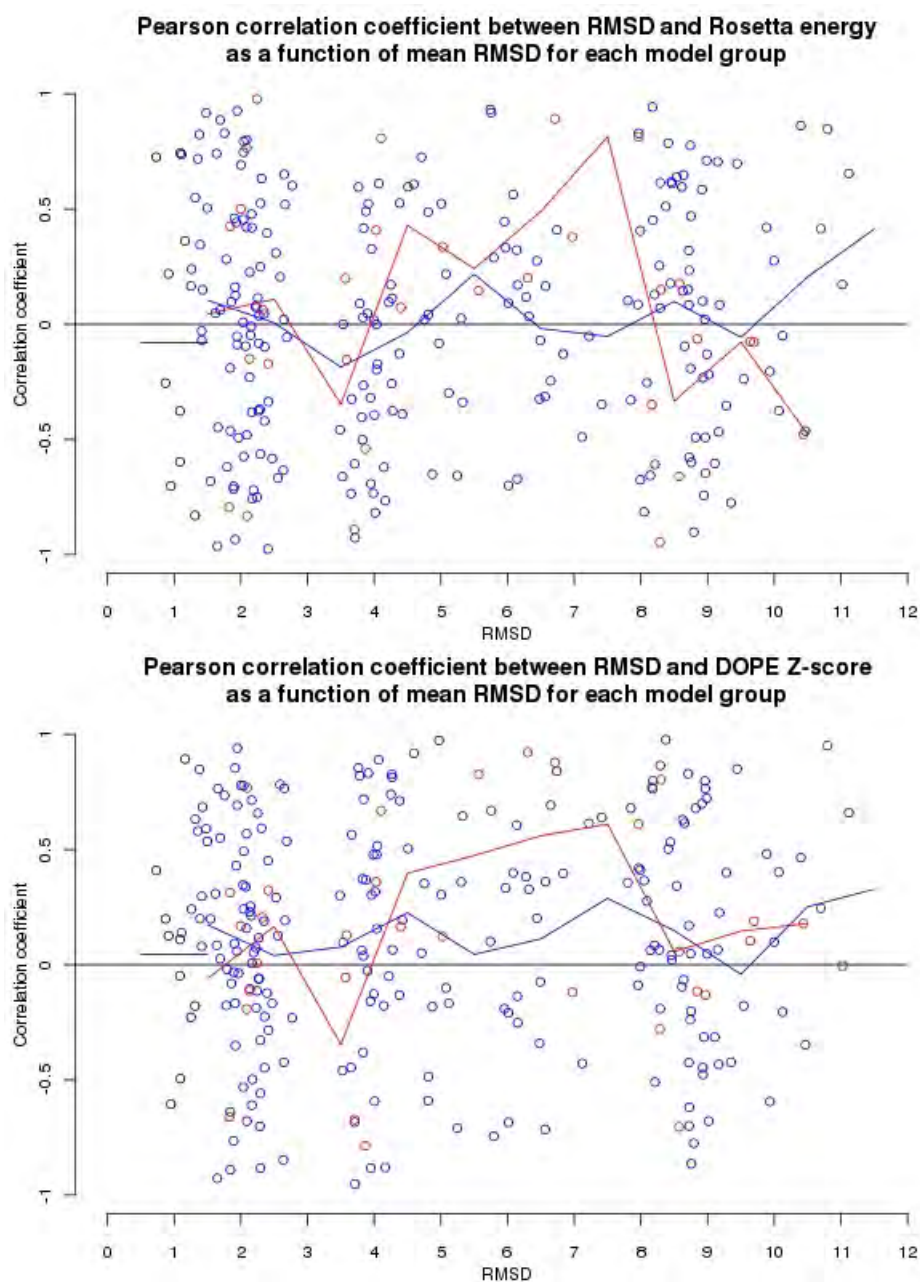


Figure 4.12 – The circles indicate the Pearson correlation coefficient between the indicated scoring functions and RMSD within each modelgroup is shown as a function of the mean RMSD of the modelgroup after minimization. The lines indicate the average value of each series over intervals of 1 Å.

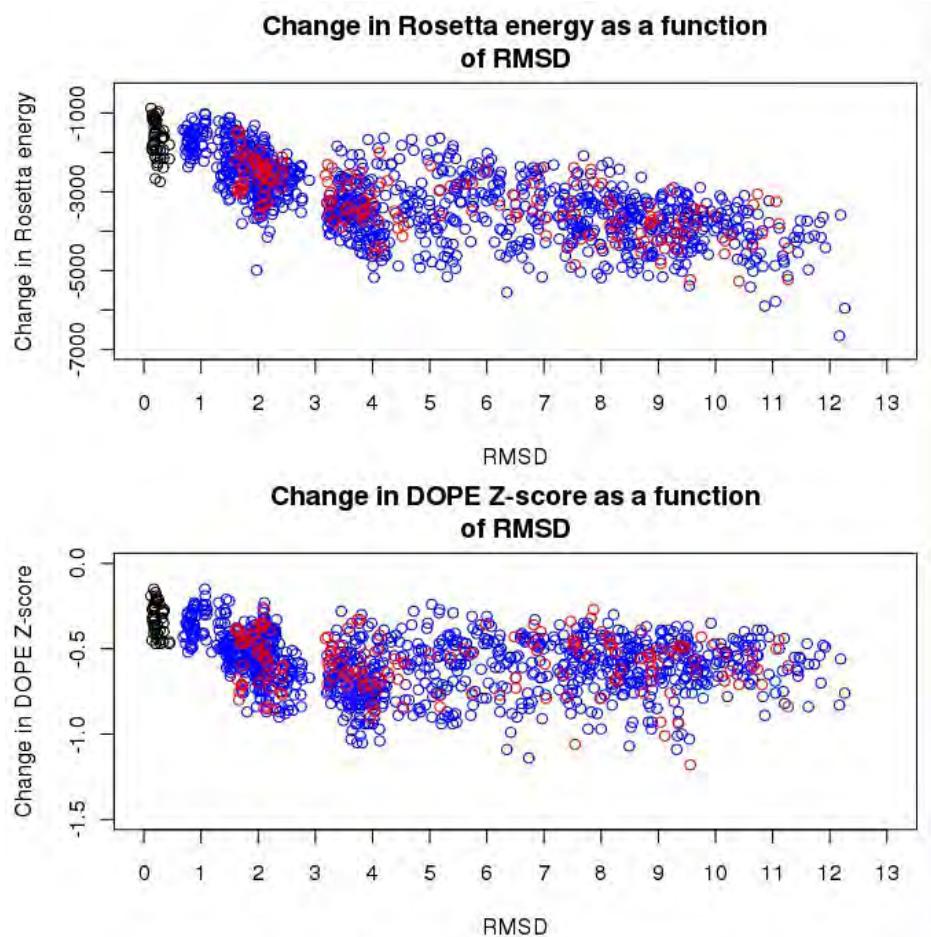


Figure 4.13 – Change in scoring functions as a function of the RMSD before minimization.

of the RMSD before minimization.

As expected, figure 4.13 shows that the worst models show the greatest improvement in score. This trend flattens off for the DOPE Z-score, which is expected as it is a normalized Z-score. A model that changes a lot during minimization is thus not as reliable as one that remains stable. Since the variation in scoring functions is significantly bigger than the variation in RMSD within a modelgroup, it is best to use the mean or median of the scoring function to represent the group than to take the value from just one of the models.

4.5 Conclusion

1325 homology models of C1 protease complexes with known experimental structures were built in order to investigate factors that can influence model accuracy. Various metrics were calculated to assess the accuracy of each model. The RMSD ranged from 0.172Å to 11.45Å.

In general, no significant difference between hybrid and single-template models, in terms of the evaluation metrics investigated, was detected. As such, hybrid templates can be used for modelling C1 complexes with cystatins whenever this enables more closely related proteins to be used for the individual proteins. This means that when modelling such a complex, the templates for the inhibitor and protease should be independently selected without regard as to what experimental structure it comes from.

It was found that both sequence identity and sequence coverage are positively correlated with accurate models. However, the correlation is not simple or linear.

The two scoring functions that were evaluated, the Rosetta energy function and the DOPE Z-score, were able to discriminate between models from different templates but could not successfully differentiate between models built on the same template. Overall, they performed very similarly. DOPE Z-score performed better in a few cases.

Many of the less accurate models ($\text{RMSD} > 4\text{\AA}$) did show improvement after minimization but these models could not be identified using the scoring functions. As such, the improvements are not practically useful. It was confirmed that less accurate models showed larger improvements in terms of their scoring functions as a result of minimization than more accurate models. The RMSD scores of accurate models did not improve and were increased in most cases.

Chapter 5

Construction of novel models

5.1 Chapter overview

In this chapter, a high-throughput modelling pipeline was used to model 110 C1 protease complexes. The chapter starts with the aim and motivation for building these particular models. The method used for modelling is then described and a list given of proteins included. The template selection and scoring function results are then given and discussed in the results and discussion section while the models themselves are available on the website at <http://rubi.ru.ac.za/cpmdb>. The chapter ends with a short conclusion.

5.2 Aim and motivation

The aim of this chapter was to accurately model complexes that are of interest due to their roles in research or disease. Cysteine proteases and inhibitors from human, malarial and plant sources were selected. The human and malarial proteins are of interest since they are important because of their role in human disease while the plant proteins, papain and taro-cystatin, are of interest due to their roles in biotechnology and food-production. For the sake of simplicity the models were constructed in an all against all fashion - combining all proteases with all inhibitors.

These models can be used for further studies investigating the complexes. Because of the large number of models, the use of statistics become possible and may enhance the knowledge gained as opposed to just building models for one specific complex of interest.

5.3 Method

5.3.1 Targets selected

The proteins selected for the study set is given in Table 5.1. Proteins were manually selected, loosely based on their importance in research and their relevance to malaria. Experimental value to the dataset were also considered. Providing diversity, providing experimental controls, and having experimental data available were considered as adding experimental value to the dataset. For example, kininogen-1 is not actually a cysteine-protease inhibitor, in spite of the sequence similarity to cystatins, so including it in the dataset serves as a form of negative control.

The selected proteases and inhibitors were combined in an all against all fashion to give the complexes to be modelled. Some of these complexes do not occur naturally but were included because it is more expedient to include them than trying to decide which have to be left out. Apart from the extra computational time required to build them, their presence does no harm.

5.3.2 Template selection

Templates for modelling were selected from the template set described in Section 2.4.2 by selecting templates for the protease and inhibitor independently from each other. For proteases, sequence coverage was used as the criteria for selecting a template while sequence identity was used for the inhibitors. The reason for the differentiation between the two is that the inhibitor gaps are situated away from the active site and thus less important. Figure 4.5 and Figure 4.6 were also considered in making this decision. The establishment of a single criteria to select templates greatly aided the automation of the modelling process.

5.3.3 Model building, energy minimization and calculation of scoring functions

After templates were selected, the modelling procedure as described in Section 2.4 was followed. This resulted in 5 models for each complex plus energy minimized versions of each of those models, all with accompanying scoring function results.

5.4 Results and discussion

5.4.1 Template selection

The result of the template selection step is shown in Table 5.2. Sequence identity and coverage values are generally high. The lowest values are 12.3% and 87.7% respectively.

Table 5.1 – Proteins selected for modelling in the study set.

Name	Organism
Proteases	
Cathepsin-B	<i>H. sapiens</i>
Cathepsin-H	<i>H. sapiens</i>
Cathepsin-L	<i>H. sapiens</i>
Cathepsin-S	<i>H. sapiens</i>
Falcipain-2	<i>P. falciparum</i>
Falcipain-2p	<i>P. falciparum</i>
Falcipain-3	<i>P. falciparum</i>
Vivapain-2	<i>P. vivax</i>
Vivapain-3	<i>P. vivax</i>
Vivapain-4	<i>P. vivax</i>
Papain	<i>C. papaya</i>
Inhibitors	
Chagasin	<i>T. cruzi</i>
Falstatin	<i>P. falciparum</i>
Cystatin-A	<i>H. sapiens</i>
Cystatin-B	<i>H. sapiens</i>
Cystatin-C	<i>H. sapiens</i>
Ovocystatin	<i>G. gallus</i>
Taro-cystatin	<i>C. esculenta</i>
Kininogen-1	<i>H. sapiens</i>
Kininogen-2	<i>H. sapiens</i>
Kininogen-3	<i>H. sapiens</i>

Table 5.2 – Results of template selection of the study set proteins. The templates chosen for each protein is indicated by its PDB ID and the sequence identity and sequence coverage are indicated in percentages. The size of each protein is given by the number of amino acid residues.

Protein name	Template	Identity	Coverage	Protein size
Proteases				
Cathepsin-B	3K9M	100.0	100.0	254
Cathepsin-H	1NB5	92.3	100.0	220
Cathepsin-L	2NQD	98.6	100.0	221
Cathepsin-S	2NQD	56.7	99.1	223
Falcipain-2	1YVB	100.0	100.0	241
Falcipain-2p	1YVB	95.9	100.0	241
Falcipain-3	1YVB	66.4	99.2	243
Vivapain-2	1YVB	62.2	99.2	243
Vivapain-3	1YVB	56.8	98.4	245
Vivapain-4	1YVB	54.6	99.6	242
Papain	3E1Z	100.0	100.0	212
Inhibitors				
Chagasin	2OUL	100.0	98.1	109
Falstatin	2OUL	12.3	97.2	110
Cystatin-A	1NB5	100.0	100.0	98
Cystatin-B	1STF	98.0	100.0	98
Cystatin-C	1YVB	45.9	92.5	120
Ovocystatin	1YVB	100.0	100.0	111
Taro-cystatin	3IMA	100.0	93.4	91
Kininogen-1	1YVB	25.9	91.5	121
Kininogen-2	1YVB	20.6	87.7	126
Kininogen-3	1YVB	28.0	87.7	126

The experimental structure 1YVB is used for many different proteins. In the protease section, it is used for all the Plasmodium proteins. Falcipain-2 is the only closely related protein to these proteases available in the template structures and occurs in 1YVB and 2OUL. As the two structures contain the same protein and hence have the same sequence identity and coverage values, the choice was arbitrarily made by the program based on the order in which they were originally entered into the program. In the case of the inhibitors, ovocystatin from 1YVB was justifiably used in many cases as it is one of only 2 type II cystatins with experimental structures in the complex form (and tarocystatin lacks a certain feature as described in Section 3.5.1).

5.4.2 Scoring function results before minimization

Table 5.3 shows the scoring function results for all the complexes of the study set before minimization. Overall, the scores are very favourable. All the DOPE Z-score values are negative (lower DOPE Z-score values indicate more accurate models) with the highest one being -0.219 for the complex with kininogen-3 and vivapain-4. The lowest value is -1.43 for the complex between cystatin-A and papain. The Rosetta energy values are all positive, ranging from 3111.4 for the complex between falstatin and vivapain-3 down to 537.6 for the complex between chagasin and cathepsin-L (lower Rosetta energy values indicate more accurate models). As Rosetta energy is not normalized to protein size, comparisons between different complexes should not be made.

5.4.3 Scoring function results after minimization

The scoring function results after minimization is shown in Table 5.4. All the Rosetta energy values are now negative and range from -827.6, for the complex between chagasin and falcipain-2, to -612.9, for the complex between falstatin and cathepsin-L. DOPE Z-scores are also much lower than before minimization and range from -1.78, for the complex between cystatin-A and cathepsin-S, to -0.814, for the complex between kininogen and VP-4.

5.5 Conclusion

Models of 110 unique complexes were built and evaluated using Rosetta energy and DOPE Z-score. These models were built in the same context and using the same techniques as the test set models and thus conclusions made regarding the test models can be extrapolated to these models. The value of these models are thus increased compared to the average homology model because of they can be compared to the models of the test set.

The models produced generally have very good DOPE Z-score and Rosetta energy values which is consistent with accurate models. Considering the sequence identity and sequence coverage at which most these models were built, it is likely that they are close to their native structures. In the case of falstatin, the models are likely inaccurate as the sequence had large segments removed to increase the sequence coverage to levels that made modelling possible.

These models represent accurate models of the respective complexes (excluding the falstatin complexes) and can be used for further studies.

Table 5.3 – Scoring function results of study set complexes before minimization. Shown is the median value of the modelgroup of each complex.

	Chagasin	Falstatin	Cyst-A	Cyst-B	Cyst-C	Ovocyst	Taro-cyst	Kini-1	Kini-2	Kini-3
Cath-B	DOPE Z	-1.09	-0.752	-1.26	-1.10	-0.789	-0.914	-0.707	-0.609	-0.608
	Rosetta	1511.7	2633.8	1660.1	1562.8	1949.4	2082.3	1866.9	1936.2	2075.7
Cath-H	DOPE Z	-1.03	-0.613	-1.08	-1.01	-0.594	-0.849	-0.582	-0.413	-0.481
	Rosetta	931.6	2061.3	1411.6	1057.4	1868.5	1122.1	1861.7	2003.2	1382.4
Cath-L	DOPE Z	-1.27	-0.868	-1.30	-1.26	-0.841	-1.06	-0.779	-0.617	-0.671
	Rosetta	537.5	1653.5	1088.1	615.9	1327.6	1038.0	1373.3	1649.0	1388.9
Cath-S	DOPE Z	-1.26	-0.888	-1.33	-1.24	-0.906	-1.06	-0.741	-0.622	-0.651
	Rosetta	1096.6	2167.0	1441.4	1060.9	1734.2	1365.8	1823.3	1924.5	1860.3
FP-2	DOPE Z	-1.03	-0.773	-1.05	-0.966	-0.650	-0.731	-0.560	-0.362	-0.439
	Rosetta	589.4	1361.2	995.8	1220.5	1438.4	1359.4	1409.5	1850.6	1510.2
FP-2p	DOPE Z	-1.01	-0.72	-0.998	-0.929	-0.650	-0.733	-0.560	-0.384	-0.437
	Rosetta	790.0	1514.3	1472.7	1238.4	1430.6	1400.6	1300.0	1654.6	1574.0
FP-3	DOPE Z	-0.777	-0.474	-0.940	-0.772	-0.429	-0.519	-0.329	-0.219	-0.313
	Rosetta	1548.9	2307.3	1354.1	1591.1	1868.3	1927.0	1984.2	2224.4	2009.0
VP-2	DOPE Z	-0.920	-0.612	-0.880	-0.868	-0.446	-0.678	-0.416	-0.339	-0.314
	Rosetta	1300.8	2013.1	1727.4	1482.2	2293.5	1688.6	1881.4	1746.1	2287.4
VP-3	DOPE Z	-0.869	-0.465	-0.808	-0.828	-0.555	-0.635	-0.349	-0.240	-0.361
	Rosetta	1389.0	3111.4	2010.6	1518.0	2035.0	1972.7	2354.8	2043.4	1926.2
VP-4	DOPE Z	-0.790	-0.481	-0.842	-0.689	-0.400	-0.554	-0.286	-0.209	-0.205
	Rosetta	1745.5	2659.0	1996.0	2189.2	2331.1	2245.9	2293.4	2472.2	2665.6
Papain	DOPE Z	-1.38	-0.955	-1.43	-1.32	-0.966	-1.09	-0.921	-0.676	-0.855
	Rosetta	610.1	1871.9	902.3	956.9	1150.2	960.0	968.7	1547.0	1024.8

Table 5.4 – Scoring function results of study set complexes after minimization. Shown is the median value of the modelgroup of each complex.

		Chagasin	Falstatin	Cyst-A	Cyst-B	Cyst-C	Ovocyst	Taro-cyst	Kini-1	Kini-2	Kini-3
Cath-B	DOPE Z	-1.40	-1.21	-1.64	-1.46	-1.26	-1.31	-1.32	-1.07	-1.04	-1.06
	Rosetta	-750.8	-616.4	-702.5	-732.9	-726.3	-722.9	-711.9	-686.3	-697.7	-688.8
Cath-H	DOPE Z	-1.38	-1.13	-1.54	-1.41	-1.16	-1.24	-1.20	-1.03	-0.948	-0.969
	Rosetta	-755.7	-650.4	-708.4	-717.9	-720.2	-724.2	-707.6	-667.6	-687.2	-697.9
Cath-L	DOPE Z	-1.53	-1.30	-1.66	-1.53	-1.30	-1.37	-1.32	-1.15	-1.06	-1.11
	Rosetta	-752.0	-612.9	-701.1	-722.0	-720.8	-724.1	-702.9	-678.1	-694.9	-691.5
Cath-S	DOPE Z	-1.59	-1.39	-1.78	-1.61	-1.40	-1.50	-1.46	-1.26	-1.20	-1.22
	Rosetta	-738.1	-663.4	-671.6	-689.4	-702.3	-690.6	-682.2	-649.2	-656.8	-671.2
FP-2	DOPE Z	-1.34	-1.22	-1.53	-1.38	-1.17	-1.24	-1.19	-1.10	-0.902	-0.981
	Rosetta	-827.6	-717.5	-759.3	-780.4	-782.8	-790.3	-777.2	-759.3	-734.4	-758.8
FP-2p	DOPE Z	-1.35	-1.12	-1.51	-1.35	-1.12	-1.21	-1.19	-1.02	-0.983	-0.977
	Rosetta	-820.9	-705.0	-767.0	-777.3	-770.6	-774.8	-773.3	-734.5	-762.7	-745.9
FP-3	DOPE Z	-1.24	-1.00	-1.41	-1.29	-1.06	-1.12	-1.09	-0.899	-0.877	-0.875
	Rosetta	-811.5	-728.3	-771.1	-779.5	-758.0	-758.3	-757.1	-729.8	-747.2	-753.5
VP-2	DOPE Z	-1.27	-1.12	-1.52	-1.36	-1.14	-1.22	-1.16	-0.987	-0.973	-0.995
	Rosetta	-790.0	-726.6	-744.1	-768.0	-754.9	-763.9	-730.6	-718.0	-739.8	-722.0
VP-3	DOPE Z	-1.29	-1.08	-1.45	-1.33	-1.16	-1.20	-1.19	-0.911	-0.956	-0.992
	Rosetta	-801.8	-711.7	-751.3	-764.0	-768.0	-752.6	-760.0	-734.0	-732.5	-735.4
VP-4	DOPE Z	-1.19	-1.06	-1.43	-1.24	-1.06	-1.10	-1.05	-0.883	-0.814	-0.904
	Rosetta	-754.0	-699.5	-709.5	-728.5	-726.1	-736.9	-699.7	-682.6	-690.0	-693.3
Papain	DOPE Z	-1.63	-1.41	-1.74	-1.62	-1.37	-1.45	-1.41	-1.21	-1.15	-1.18
	Rosetta	-758.4	-648.6	-696.1	-711.3	-714.4	-710.6	-701.6	-687.4	-687.4	-697.4

Chapter 6

Creation of website

6.1 Chapter overview

This chapter discusses the website that accompanies this thesis. The reason for providing the website is explained at the start of the chapter, more detail is provided on the technical aspects of the development in the method section followed by a description of the website in the results and discussion section. The chapter is finished off with a conclusion about the website.

6.2 Aim and motivation

The project produced over a thousand models and a large amount of useful information about the accuracy of the modelling process itself and the performance of two specific scoring functions. Most of this data is useful in itself and can be analyzed further to gain insight into the homology modelling process, the scoring functions examined and the relationships between the different proteins. This information can only be analyzed electronically, due to the nature and amount, and as such is more useful in an electronic format. To encourage further use of the data it was decided to make it publicly available on a website. Such a website was developed and called CPmDB (Cysteine Protease model DataBase). The aim of the website is to provide a user-friendly interface to access the raw data.

6.3 Method

It was decided to use the Django framework to develop the website. Django is a rapid web-development-framework written in the Python programming language that helps developers to quickly create dynamic websites and enforces the use of good programming practices.

Django uses the Model-Controller-Viewer paradigm to separate the different aspects of a dynamic website into 3 distinct parts.

Model Responsible for storage and retrieval of raw data.

Controller Responsible for processing data and any complex operations that form part of the website functionality.

Viewer Presentation of data in a suitable layout.

For the model layer a MySQL database is used to store the data. Django has its own database interface to interact with the database. The reason why the original python modules that interact with the project data could not be used directly, is that loading them takes too long. The MySQL database allows Django to load only the required data, and improves the performance of the website significantly by taking advantage of modern relational database technology. Python functions were written to load the data from the python modules, used in the rest of the project, into the database. The database was loaded with all of the data and molecular structures used in the rest of the project by running these functions.

The controller is a python module where certain functions are linked to specific URLs via regular expressions (regular expressions are a common way to define patterns in character sequences). These functions must return the formatted page that the user receives in his/her browser. The functions and procedures of the controller performs any processing of data that is necessary to generate the page requested by the user. While the controller can produce the formatted page directly, it is more common and better programming practice to leave formatting the viewer part of the program.

In Django, the viewer takes the form of a template language and python modules that facilitate the loading and rendering of templates. Templates are text files with special tags for specifying parts of the page that can be changed. These parts have to be correctly specified each time the template is used. It is the job of the controller to provide data to the viewer so these fields in the template can be correctly populated. The idea is to provide all the formatting and layout information in invariable part of the template. Some logic and control functions can be performed in the template as well to accommodate cases where is uncertainty about the nature of the data that needs to be displayed. The logic and control functions allow the template to adapt to the data it receives from the controller.

6.4 Results and discussion

6.4.1 General

The website was successfully completed and is available at <http://rubi.ru.ac.za/cpmdb>.

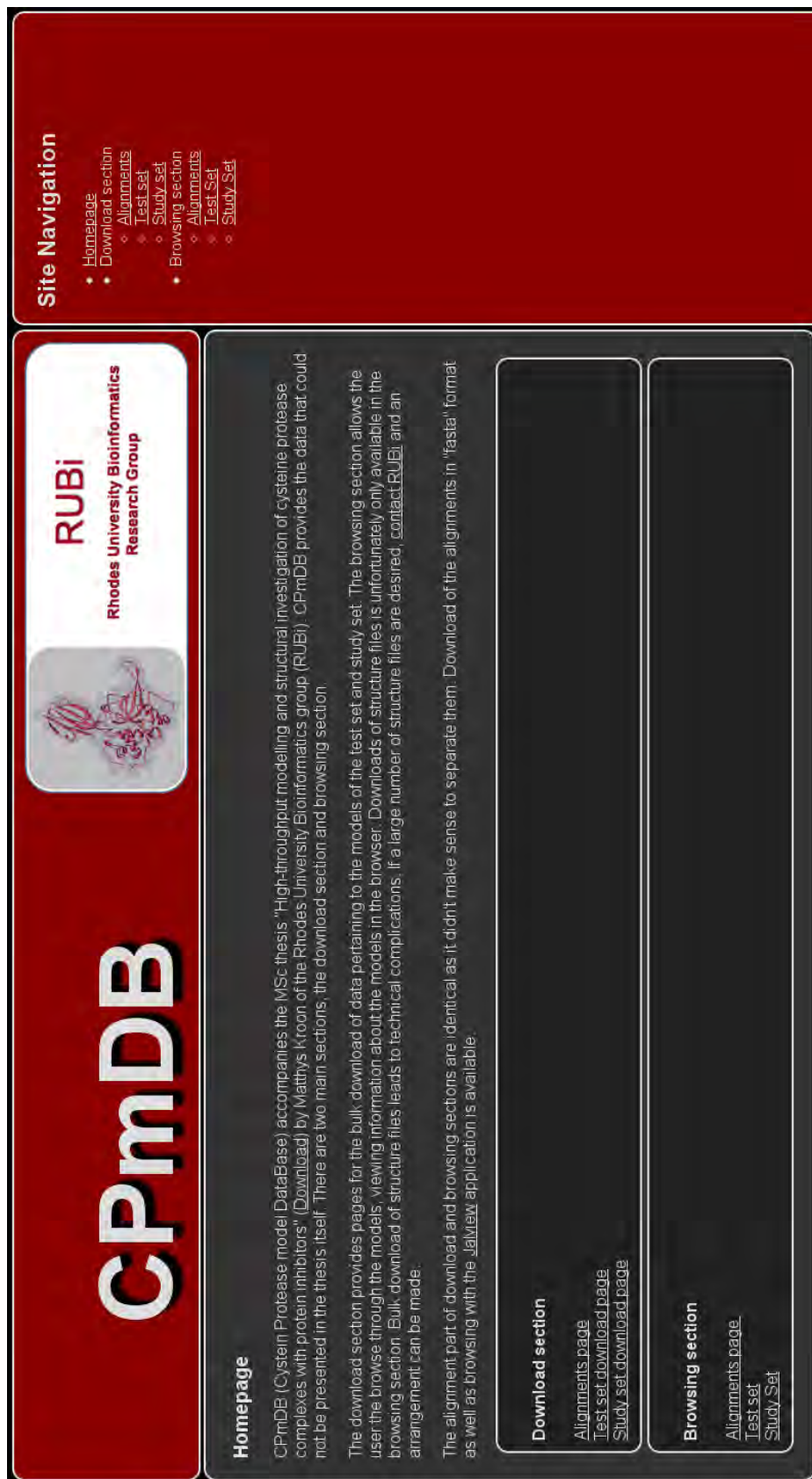


Figure 6.1 – Screenshot of the website homepage

Figure 6.1 shows a screenshot of the homepage. A navigation panel on the right provides quick access to the different sections. Content is shown in the main display window which fills up most of the space.

The content is divided into two main sections, the download section and browsing section. The download section allows users to download bulk data pertaining to the project. It has separate pages for the alignment, test set data and study set data. The browsing section allows users to explore the data by following links through a hierarchy, representing the project. For alignments, the browsing section and download section is the same.

6.4.2 Alignment section

The alignment section consists of page showing a list of the alignments used and a page for each of the alignments. The user accesses the individual alignment pages by clicking the items in the list. On the individual alignment page a list of all the proteins in the alignment is given. The alignment itself can be browsed using the Jalview (<http://www.jalview.org/>) application. The alignment can also be downloaded in fasta format. These pages allow both a detailed view of the individual alignments as well as the ability to download the information. The alignment link in the download section and browsing section points to the same page.

6.4.3 Download section

The download section provides filters with which to select subsets of the data to download. Such filters are given for the study set and test separately. The filters act together in such a fashion that the complex must satisfy the condition of all the filters to be included in the results. A choice is given whether the user wants to download median values for each modelgroup or the individual results of each model. Upon setting the filters, the user is presented with a list of complexes that satisfy the criteria. The user can then choose to download the information for the selected complexes or to change the filter settings. When the user chooses to download the results, a CSV (comma separated values) file is generated and sent to the user. The CSV file can be opened in most spreadsheet programs.

6.4.4 Browsing section

The browsing section allows the user to peruse the data. In addition to the alignment section described already there are sections for the test set and study set. The link for each set is organized in a different way. The test set uses a hierarchy with two levels. The highest level is target complex and the second is modelgroup. The study set uses only one level which is modelgroup. In both study set and test set, a page is available for each modelgroup on which all the data for each model and the structure files of each model is available.

6.5 Conclusion

The website was successfully constructed and all the data produced in the project is available there. It is hoped that other researchers will make use of the data, the proper analysis of which was beyond the scope of this project. The interface is relatively simple. Initially, access to the website is be restricted by requiring a username and password to view the content. This is a temporary restriction that will be lifted as soon as documents, currently in writing, can be published. Access to data is an important cornerstone of science in general and especially bioinformatics. It is hoped that giving free and unlimited access to research data will become the standard practice in future.

Chapter 7

Conclusion

7.1 Chapter Overview

In this project, the high-throughput modelling of protein complexes containing C1 cysteine proteases and protein inhibitors were studied. To this end the following actions were carried out:

1. The structural variation between the protein sub-families, and conformational changes upon complex formation were examined.
2. Test models were built, to evaluate modelling accuracy, and accuracy metrics for these models were calculated.
3. Hybrid templates were introduced and a comparison between hybrid models, and simple models were carried out.
4. The effect of energy minimization using Rosetta all-atom energy minimization was examined.
5. Two metrics, the Rosetta Energy and the DOPE Z-score, were examined for their ability to discriminate between models of varying accuracy.
6. Several complexes of unknown structure were modelled using information from the preceding studies to select the best templates.
7. A website was created for distribution of data pertaining to the project.

This chapter tries to make conclusions based on the results of these actions as presented in the previous chapters. Some conclusions draw from more than one of the chapters and so they are not ordered according to chapter. A section is given for each logically separate conclusion, followed by a closing paragraph.

7.2 Conformational change can occur during complex formation between C1 proteases and protein inhibitors

The results of chapter 3 showed that some of the proteins examined undergo conformational change upon binding. The most significant change is arguably the movement of the cathepsin-B occluding loop to accommodate the inhibitors in the active site cleft. Cathepsin-B has a clearly favoured conformation for the occluding loop as is shown by the structure 1GMV where all three conformations of monomeric cathepsin-B show the same conformation for the occluding loop. Upon complex formation the occluding loop is pushed away from this favoured position.

Both cystatin subfamilies also undergo conformational changes upon binding. Most notably, the N-terminal region binds to the protease in a stable conformation while being highly mobile in the monomeric state. Differences between the cystatin-A conformations in 1NB5 (with cathepsin-H) and 3K9M (with cathepsin-B) show that cystatin-A adapts its conformation to fit the protease its binding to.

Chagasin shows more similar conformations between the bound and monomeric form than cystatin. There is however one interesting structure where chagasin shows an interaction with papain not present in any of the other complexes where chagasin undergoes a slight conformational change (3E1Z).

Papain does not undergo conformational change upon binding to cystatin or chagasin, it shows almost perfect similarity between 4 different crystal structures which includes 3 complexes and one monomeric structure.

The main implication here with regards to the modelling of complexes is that, with the exception of cathepsin-L like proteases, templates for homology modelling must be taken from complexes and not from monomeric structures.

7.3 Model accuracy could be estimated sensibly by measuring similarity to experimental structures

After comparing different experimental structures to each other, it was concluded that 0.38Å can be used as a rough estimate for the RMSD below which the relationship between similarity to an experimental structure, and similarity to a native protein structure, breaks down. Above this value, comparison of models to experimental structures is a good method of estimating accuracy. Below this value, the models are so close to the native structure that the errors in experimental structures are of the same magnitude as the errors in the models. Of course this value is only a very rough estimate and the actual number might be higher or lower and vary from protein to protein. The best models constructed in the test set had RMSD values of around 0.7Å and these were exceptional models where both protease and inhibitor were modelled at very high sequence similarity. Most models are significantly less accurate than

this and thus the small difference between the actual native structure and the experimental structures are not a real factor in the measurement of the model accuracy.

7.4 There was no significant difference between the accuracy of hybrid and simple models

By comparing hybrid and simple models, from a subset of the test models, directly, it was shown that they are of equal quality. This was done by selecting a subset of test models so that the hybrid and simple models in the subset had the same distribution of sequence similarity. Each of the major evaluation metrics were then compared between the hybrid and simple models using the Mann-Whitney U-test. The result showed that the differences between the two classes could be explained by random chance. Additionally, histograms of the evaluation metrics showed similar distributions for the two groups. The differences between the means of the two groups were also very small when compared to the values that were being measured.

Since hybrid templates and simple templates produce models of the same quality, templates for the individual components of the complexes (proteases and inhibitors) can be selected independently from separate experimental structures. This means that, in most cases, templates with higher sequence similarity can be used. Further evidence that hybrid templates are useful in building more accurate models is the fact that, for each of the nine targets in the test set, the best model was based on a hybrid template in all nine cases. This proves that, even if there were a slight cost to using hybrid templates, the benefit of using templates with higher sequence similarity more than makes up for that cost.

7.5 Scoring functions provide useful information in certain circumstances

It was found that both scoring functions examined, Rosetta energy and DOPE Z-score, was able to identify the more accurate modelgroups. Both functions showed positive correlation with RMSD, with DOPE Z-score showing slightly better performance. If there is uncertainty about which template to use for a particular target, two or more potential templates can be used to generate a few models each. Those models can then be evaluated with the scoring functions to indicate which models are likely to be more accurate. The template that produces the lowest scoring models can then be used further.

When it comes to selecting the best models out of a modelgroup, both functions generally fail. This conclusion applies only to the specific method followed in this project. With other modelling protocols the result might be different. However, the protocol followed here is relatively common and there is reason to think that the result applies more widely than just

for cysteine protease complexes. In the case of the self models, both functions were able to make better-than-random selections of accurate models. What difference between the self models and other models allows this is unclear. The only obvious difference is that the self models are all highly accurate, with RMSD under 0.5Å. There were not enough models of the hybrid and simple class in this accuracy range to get a sensible indication of the performance of the scoring functions.

After energy minimization with the Rosetta 3.1 relax protocol, the ability of DOPE Z-score to detect the more accurate models in a modelgroup increased very slightly across the entire range of accuracy, although the correlation is still barely better than random.

7.6 Rosetta energy minimization improves the accuracy of some models

For most of the very accurate models the energy minimization led to a decrease in accuracy in terms of all evaluation metrics except fraction of native contacts. For less accurate models a fraction of them were improved by the minimization procedure. The fraction of models that were improved increases for less accurate models. One very interesting observation was that the number of residue-residue contacts as defined for F-nat and F-non-nat increased significantly for almost all models. This means that the energy minimization leads to tighter packing of the interface residues and brings the two proteins closer to each other.

While there were improvements to many models, the practical use of the energy minimization step is limited as the scoring functions are not able to identify the models that were improved. If it was possible to identify the improved models, the utility of the minimization procedure would be greatly improved.

7.7 Models for novel complexes were built and made available in a website

Models for a hundred and ten complexes were built using the same modelling protocol as for the test set. They were evaluated using the same scoring functions and subjected to energy minimization just like the test set models. Considering the sequence similarity between the templates and the targets, the models are likely to be of high quality. The exception to this statement is the models involving falstatin. The sequence similarity between falstatin and chagasin was quite low and showed large insert, one of them very close to the interface. Given this, it is unlikely that the falstatin models are very accurate although the relatively low energies for some of the falstatin models do suggest that the sequence fits the structure well.

All the models of both sets as well as their scoring function results and evaluation metrics in the case of the test models were made available online at <http://rubi.ru.ac.za/cpmdb>. This was done so that other researchers can further analyze the data or use the models in studies.

7.8 Closing

The ultimate aim of the project was to model a large number of novel cysteine protease complexes with very exact estimates of their accuracy. In this regard, the project fell short. However, many important new insights were drawn from the investigations undertaken. The most significant finding was that the hybrid templates produce models of similar quality to the simple templates. How extensible this is to other protein complexes remains to be seen. A lot of the software developed in the project is also re-usable for future research involving protein structures and especially research involving high-throughput modelling.

Bibliography

- Abrahamson, M. 1988. Human cysteine proteinase inhibitors. Isolation, physiological importance, inhibitory mechanism, gene structure and relation to hereditary cerebral hemorrhage. *Scandinavian Journal of Clinical & Laboratory Investigation*, **191**, 21–31.
- Alcala-Canto, Y, Ibarra-Velarde, F, Sumano-Lopez, H, Gracia-Mora, J, & Alberti-Navarro, A. 2007. Effect of a cysteine protease inhibitor on *Fasciola hepatica* (liver fluke) fecundity, egg viability, parasite burden, and size in experimentally infected sheep. *Parasitology Research*, **100**(3), 461–465.
- Altschul, S F, Madden, T L, Schaffer, A A, Zhang, J, Zhang, Z, Miller, W, & Lipman, D J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**(17), 3389–3402.
- Andricopulo, AD, Salum, LB, & Abraham, DJ. 2009. Structure-Based Drug Design Strategies in Medicinal Chemistry. *Current Topics in Medicinal Chemistry*, **9**(9), 771–790.
- Anfinsen, CB, & Haber, E. 1961. Studies on the reduction and re-formation of protein disulfide bonds. *Journal of Biological Chemistry*, **236**(5), 1361–1363.
- Apweiler, R, Martin, MJ, O'Donovan, C, Magrane, M, Alam-Faruque, Y, Antunes, R, Barrell, D, Bely, B, Bingley, M, Binns, D, & Others. 2010. The universal protein resource (UniProt) in 2010. *Nucleic Acids Research*, **38**, 142–148.
- Ashie, INa, Sorensen, TL, & Nielsen, PM. 2002. Effects of Papain and a Microbial Enzyme on Meat Proteins and Beef Tenderness. *Journal of Food Science*, **67**(6), 2138–2142.
- Barrett, AJ. 1981. Cystatin, the egg white inhibitor of cysteine proteinases. *Methods in Enzymology*, **80**, 771–778.
- Ben-David, M, Noivirt-Brik, O, Paz, A, Prilusky, J, Sussman, JL, & Levy, Y. 2009. Assessment of CASP8 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, **77**(S9), 50–65.
- Benkert, P, Biasini, M, & Schwede, T. 2011. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, **27**(3), 343–350.

- Berman, HM, Westbrook, J, Feng, Z, Gilliland, G, Bhat, TN, Weissig, Helge, Shindyalov, IN, & Bourne, PE. 2000. The Protein Data Bank. *Nucleic Acids Research*, **28**(1), 235–242.
- Bode, W, Engh, R, Musil, D, Thiele, U, Huber, R, Karshikov, A, Brzin, J, Kos, J, & Turk, V. 1988. The 2.0 Å X-ray crystal structure of chicken egg white cystatin and its possible mode of interaction with cysteine proteinases. *The EMBO Journal*, **7**(8), 2593–2599.
- Caffrey, CR, Salter, JP, Lucas, KD, Khiem, D, Hsieh, I, Lim, K, Ruppel, A, McKerrow, JH, & Sajid, M. 2002. SmCB2, a novel tegumental cathepsin B from adult *Schistosoma mansoni*. *Molecular and Biochemical Parasitology*, **121**(1), 49–61.
- Cazzulo, JJ, Stoka, V, & Turk, V. 1997. Cruzipain, the major cysteine proteinase from the protozoan parasite *Trypanosoma cruzi*. *Journal of Biological Chemistry*, **378**(1), 1–10.
- Chothia, C, & Lesk, AM. 1986. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, **5**(4), 823–826.
- Da Silva, AAF, De Carvalho Vieira, L, Krieger, MA, Goldenberg, S, Zanchin, NIT, & Guimarães, BG. 2007. Crystal structure of chagasin, the endogenous cysteine-protease inhibitor from *Trypanosoma cruzi*. *Journal of Structural Biology*, **157**(2), 416–423.
- Dalton, JP, Neill, SO, Stack, C, Collins, P, Walshe, A, Sekiya, M, Doyle, S, Mulcahy, G, Hoyle, D, Khaznadji, E, Moiré, N, Brennan, G, Mousley, A, Kreshchenko, N, Maule, AG, & Donnelly, SM. 2003. *Fasciola hepatica* cathepsin L-like proteases: biology, function, and potential in the development of first generation liver fluke vaccines. *International Journal for Parasitology*, **33**(11), 1173–1181.
- DeLano, WL. 2002. *The PyMOL Molecular Graphics System*.
- Dos Santos Monteiro, AC, De Oliveira Neto, OB, Del Sarto, RP, De Magalhães, MTQ, Lima, JN, Lacerda, AF, Oliveira, RS, Scharfstein, J, Da Silva, MCM, Valencia, JWA, Jiménez, AV, & Grossi-de Sa, MF. 2008. A recombinant form of chagasin from *Trypanosoma cruzi*: inhibitory activity on insect cysteine proteinases. *Pest Management Science*, **64**(7), 755–60.
- Dowd, A J, Smith, A M, McGonigle, S, & Dalton, J P. 1994. Purification and characterisation of a second cathepsin L proteinase secreted by the parasitic trematode *Fasciola hepatica*. *European Journal of Biochemistry*, **223**(1), 91–98.
- Drenth, J, Jansonius, JN, Koekoek, R, Marrink, J, Munnik, J, & Wolthers, BG. 1962. The crystal structure of papain. Two-dimensional fourier syntheses. *Journal of Molecular Biology*, **5**(4), 398–407.
- Dvorák, J, Mashiyama, ST, Sajid, M, Braschi, S, Delcroix, M, Schneider, EL, McKerrow, WH, Bahgat, M, Hansell, E, Babbitt, PC, Craik, CS, McKerrow, JH, & Caffrey, CR. 2009. SmCL3, a gastrodermal cysteine protease of the human blood fluke *Schistosoma mansoni*. *PLoS Neglected Tropical Diseases*, **3**(6), e449.

- Ettari, R, Bova, F, Zappalà, M, Grasso, S, & Micale, N. 2010. Falcipain-2 inhibitors. *Medicinal Research Reviews*, **30**(1), 136–167.
- Finn, RD, Mistry, J, Tate, J, Coghill, P, Heger, A, Pollington, JE, Gavin, O L, Gunasekaran, P, Ceric, G, Forslund, K, Holm, L, Sonnhammer, ELL, Eddy, SR, & Bateman, A. 2010. The Pfam protein families database. *Nucleic Acids Research*, **38**(Database issue), D211–D222.
- Fossum, K, & Whitaker, JR. 1968. Ficin and papain inhibitor from chicken egg white. *Archives of Biochemistry and Biophysics*, **125**(1), 367–375.
- Freitas, MAR, Fernandes, HC, Calixto, VC, Martins, AS, Silva, EF, Pesquero, JL, & Gomes, MA. 2009. *Entamoeba histolytica*: cysteine proteinase activity and virulence. Focus on cysteine proteinase 5 expression levels. *Experimental Parasitology*, **122**(4), 306–9.
- Greer, J. 1981. Comparative model-building of the mammalian serine proteases. *Journal of Molecular Biology*, **153**(4), 1027–1042.
- Hall, A, Hakansson, K, Mason, RW, Grubb, A, & Abrahamson, M. 1995. Structural Basis for the Biological Specificity of Cystatin C. *Journal of Biological Chemistry*, **270**(10), 5115–5121.
- Hanspal, M, Dua, M, Takakuwa, Y, Chishti, AH, & Mizuno, A. 2002. Plasmodium falciparum cysteine protease falcipain-2 cleaves erythrocyte membrane skeletal proteins at late stages of parasite development. *Blood*, **100**, 1048–1054.
- Hartmann, S, & Lucius, R. 2003. Modulation of host immune responses by nematode cystatins. *International Journal for Parasitology*, **33**(11), 1291–1302.
- Hayton, K, & Su, X. 2004. Genetic and biochemical aspects of drug resistance in malaria parasites. *Current Drug Targets - Infectious Disorders*, **4**(1), 1–10.
- Hillisch, A, Pineda, LF, & Hilgenfed, R. 2004. Utility of homology models in the drug discovery. *Drug Discovery Today*, **9**, 659–669.
- Hwang, H, Mintseris, J, Pierce, B, Janin, J, & Weng, Z. 2008. Protein-protein docking benchmark version 3.0. *Proteins*, **73**(3), 705.
- Janin, J, Henrick, K, Moult, J, Eyck, LT, Sternberg, MJE, Vajda, S, Vakser, I, & Wodak, SJ. 2003. CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*, **52**(1), 2–9.
- Jeong, J, Kumar, A, Hanada, T, Seo, P, Li, X, Hanspal, M, & Chishti, AH. 2006. Cloning and characterization of Plasmodium falciparum cysteine protease, falcipain-2B. *Blood Cells, Molecules & Diseases*, **36**(3), 429–35.
- Jimin, P. 2008. Multiple protein sequence alignment. *Current Opinion in Structural Biology*, **18**(3), 382–386.

- Jones, S, & Thornton, JM. 1996. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(1), 13–20.
- Kendrew, JC, Bodo, G, Dintzis, HM, Parrish, RG, Wyckoff, H, Phillips, DC, & Others. 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, **181**(4610), 662–666.
- Kryshtafovych, A, & Fidelis, K. 2009. Protein structure prediction and model quality assessment. *Drug Discovery Today*, **14**(7-8), 386–393.
- Kryshtafovych, A, Fidelis, K, & Tramontano, A. 2011. Evaluation of model quality predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics*, **79**(S10).
- Kundrotas, PJ, & Alexov, E. 2006. Predicting 3D structures of transient protein-protein complexes by homology. *Biochimica et Biophysica Acta*, **1764**(9), 1498–511.
- Kundrotas, PJ, & Vakser, IA. 2010. Accuracy of Protein-Protein Binding Sites in High-Throughput Template-Based Modeling. *PLoS Computational Biology*, **6**(4), e1000727.
- Lensink, MF, & Wodak, SJ. 2010. Docking and scoring protein interactions: CAPRI 2009. *Proteins: Structure, Function, and Bioinformatics*, **78**(15), 3073–3084.
- Ljunggren, A, Redzynia, I, Alvarez-Fernandez, M, Abrahamson, M, Mort, JS, Krupa, JC, Jaskolski, M, & Bujacz, G. 2007. Crystal structure of the parasite protease inhibitor chagasin in complex with a host target cysteine protease. *Journal of Molecular Biology*, **371**(1), 137–153.
- Lopes, MC, Mascarini, RC, Da Silva, BMCG, Florio, FM, & Basting, RT. 2007. Effect of a papain-based gel for chemomechanical caries removal on dentin shear bond strength. *Journal of Dentistry for Children*, **74**(2), 93–97.
- Mahmoudzadeh-Niknam, H, & McKerrow, JH. 2004. Leishmania tropica: cysteine proteases are essential for growth and pathogenicity. *Experimental Parasitology*, **106**(3-4), 158–163.
- Mann, HB, & Whitney, DR. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, **18**(1), 50–60.
- Martin, SHC. 1885. Papain-Digestion. *The Journal of Physiology*, **5**(4-6), 213.
- Martinez, M, & Diaz, I. 2008. The origin and evolution of plant cystatins and their target cysteine proteinases indicate a complex functional relationship. *BMC Evolutionary Biology*, **8**(Jan.), 198.
- McKerrow, JH, Rosenthal, PJ, Swenerton, R, & Doyle, P. 2008. Development of protease inhibitors for protozoan infections. *Current Opinion in Infectious Diseases*, **21**(6), 668–672.

- McLachlan, AD. 1982. Rapid comparison of protein structures. *Acta Crystallographica*, **38**(6), 871–873.
- Méndez, R, Leplae, R, De Maria, L, & Wodak, SJ. 2003. Assessment of blind predictions of protein-protein interactions: Current status of docking methods. *Proteins: Structure, Functions and Bioinformatics*, **52**(1), 51–67.
- Misura, KMS, Chivian, D, Rohl, CA, Kim, DE, & Baker, D. 2006. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(14), 5361–5366.
- Moncada, D, Keller, K, & Chadee, K. 2003. Entamoeba histolytica cysteine proteinases disrupt the polymeric structure of colonic mucin and alter its protective function. *Infection and Immunity*, **71**(2), 838–44.
- Monteiro, ACS, Abrahamson, M, Lima, APCA, Vannier-Santos, MA, & Scharfstein, J. 2001. Identification, characterization and localization of chagasin, a tight-binding cysteine protease inhibitor in Trypanosoma cruzi. *Journal of Cell Science*, **114**(Pt 21), 3933–42.
- Mottram, JC, Coombs, GH, & Alexander, J. 2004. Cysteine peptidases as virulence factors of Leishmania. *Current Opinion in Microbiology*, **7**(4), 375–81.
- Moult, J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, **15**(3), 285–289.
- Moult, J, Pedersen, JT, Judson, R, & Fidelis, K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, **23**(3), ii–v.
- Moult, J, Fidelis, K, Kryshtafovych, A, Rost, B, Hubbard, T, & Tramontano, A. 2007. Critical assessment of methods of protein structure prediction—Round VII. *Proteins: Structure, Function, and Bioinformatics*, **69**(Suppl 8), 3.
- Mussap, M, & Plebani, M. 2004. Biochemistry and clinical role of human cystatin C. *Critical Review in Clinical Laboratory Sciences*, **41**(5-6), 467–550.
- Na, B, Kim, T, Rosenthal, PJ, Lee, J, & Kong, Y. 2004a. Evaluation of cysteine proteases of Plasmodium vivax as antimalarial drug targets: sequence analysis and sensitivity to cysteine protease inhibitors. *Parasitology Research*, **94**(4), 312–317.
- Na, B, Shenai, BR, Sijwali, PS, Choe, Y, Pandey, KC, Singh, A, Craik, CS, & Rosenthal, PJ. 2004b. Identification and biochemical characterization of vivapains, cysteine proteases of the malaria parasite Plasmodium vivax. *The Biochemical Journal*, **378**(Pt 2), 529–38.
- Notredame, C, Higgins, DG, & Heringa, J. 2000. T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–217.

- Pandey, KC, Singh, N, Arastu-Kapur, S, Bogyo, M, & Rosenthal, PJ. 2006. Falstatin, a cysteine protease inhibitor of *Plasmodium falciparum*, facilitates erythrocyte invasion. *PLoS Pathogens*, **2**(11), e117.
- Pei, J, Kim, B, & Grishin, NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Research*, **36**(7), 2295–300.
- Pennacchio, LA, Lehesjoki, A, Stone, NE, Willour, VL, Virtaneva, K, Miao, J, D’Amato, E, Ramirez, L, Faham, M, Koskiniemi, M, Warrington, JA, Norio, R, De la Chapelle, A, Cox, DR, & Myers, RM. 1996. Mutations in the Gene Encoding Cystatin B in Progressive Myoclonus Epilepsy (EPM1). *Science*, **271**(5256), 1731–1734.
- Petrey, D, Xiang, Z, Tang, CL, Xie, L, Gimpelev, M, Mitros, T, Soto, CS, Goldsmith-Fischman, S, Kernytsky, A, Schlessinger, A, Koh, IYY, Alexov, E, & Honig, B. 2003. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins: Structure, Function, and Bioinformatics*, **53**(S6), 430–435.
- Price, RN, Tjitra, E, Guerra, CA, Yeung, S, White, NJ, & Anstey, NM. 2007. Vivax malaria: neglected and not benign. *The American Journal of Tropical Medicine and Hygiene*, **77**(6 Suppl), 79–87.
- Pusey, ML, Liu, ZJ, Tempel, W, Praissman, J, Lin, D, Wang, BC, Gavira, JA, & Ng, JD. 2005. Life in the fast lane for protein crystallization and X-ray crystallography. *Progress in Biophysics and Molecular Biology*, **88**(3), 359–386.
- Ranjit, N, Zhan, B, Hamilton, B, Stenzel, D, Lowther, J, Pearson, M, Gorman, J, Hotez, P, & Loukas, A. 2009. Proteolytic degradation of hemoglobin in the intestine of the human hookworm *Necator americanus*. *The Journal of Infectious Diseases*, **199**(6), 904–12.
- Rawlings, ND, & Barrett, AJ. 1994. Families of cysteine peptidases. *Methods in Enzymology*, **244**, 461.
- Rawlings, ND, Tolle, DP, & Barrett, AJ. 2004. Evolutionary families of peptidase inhibitors. *Biochemical Journal*, **378**(Pt 3), 705–716.
- Rawlings, ND, Barrett, AJ, & Bateman, A. 2010. MEROPS: the peptidase database. *Nucleic Acids Research*, **38**(Database issue), D227–33.
- Read, RJ, & Chavali, G. 2007. Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins: Structure, Function and Bioinformatics*, **69** Suppl 8(Suppl 8), 27–37.
- Redzynia, I, Ljunggren, A, Abrahamson, M, Mort, JS, Krupa, JC, Jaskolski, M, & Bujacz, G. 2008. Displacement of the occluding loop by the parasite protein, chagasin, results in efficient inhibition of human cathepsin B. *The Journal of Biological Chemistry*, **283**(33), 22815–25.

- Rennenberg, A, Lehmann, C, Heitmann, A, Witt, T, Hansen, G, Nagarajan, K, Deschermeier, C, Turk, V, Hilgenfeld, R, & Heussler, VT. 2010. Exoerythrocytic Plasmodium parasites secrete a cysteine protease inhibitor involved in sporozoite invasion and capable of blocking cell death of host hepatocytes. *PLoS Pathogens*, **6**(3), e1000825.
- Rigden, DJ, & Mosolov, VV. 2002. Sequence conservation in the chagasin family suggests a common trend in cysteine proteinase binding by unrelated protein inhibitors. *Protein Science*, **11**, 1971–1977.
- Rigden, DJ, Monteiro, A, & Grossi de Sá, MF. 2001. The protease inhibitor chagasin of *Trypanosoma cruzi* adopts an immunoglobulin-type fold and may have arisen by horizontal gene transfer. *Federation of European Biochemical Societies Letters*, **504**(1-2), 41–44.
- Rosenthal, PJ. 2004. Cysteine proteases of malaria parasites. *International Journal for Parasitology*, **34**, 1489–1499.
- Rosenthal, PJ, Lee, GK, & Smith, RE. 1993. Inhibition of a *Plasmodium vinckei* cysteine proteinase cures murine malaria. *The Journal of Clinical Investigation*, **91**(3), 1052–6.
- Sajid, M, & McKerrow, J H. 2002. Cysteine proteases of parasitic organisms. *Molecular and Biochemical Parasitology*, **120**(1), 1–21.
- Sali, A, & Blundell, TL. 1993. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, **234**(3), 779–815.
- Sali, A, & Shen, M. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Science: A Publication of the Protein Society*, **15**(11), 2507–2524.
- Salmon, D, Do Aido-Machado, R., Diehl, A, Leidert, M, Schmetzer, O, & Others. 2006. Solution structure and backbone dynamics of the *Trypanosoma cruzi* cysteine protease inhibitor chagasin. *Journal of Molecular Biology*, **357**(5), 1511–1521.
- Sanchez, R, & Sali, A. 1997. Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology*, **7**(2), 206–214.
- Santos, CC, Scharfstein, J, & de A. Lima, APC. 2006. Role of chagasin-like inhibitors as endogenous regulators of cysteine proteases in parasitic protozoa. *Parasitology Research*, **99**(4), 323–324.
- Shatsky, M, Nussinov, R, & Wolfson, HJ. 2004. A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics*, **56**(1), 143–156.
- Shaw, DE, Deneroff, MM, Dror, RO, Kuskin, JS, Larson, RH, Salmon, JK, Young, C, Batson, B, Bowers, KJ, Chao, JC, Eastwood, MP, Gagliardo, J, Grossman, JP, Ho, CR, Ierardi, DJ, Kolossváry, I, Klepeis, JL, Layman, T, McLeavey, C, Moraes, MA, Mueller, R, Priest, EC, Shan, Y, Spengler, J, Theobald, M, Towles, B, & Wang, SC. 2007. Anton, a

- special-purpose machine for molecular dynamics simulation. *ACM SIGARCH Computer Architecture News*, **35**(2), 1–12.
- Shenai, BR, Sijwali, PS, Singh, A, & Rosenthal, PJ. 2000. Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum*. *The Journal of Biological Chemistry*, **275**(37), 29000–10.
- Shindo, T, & Van der Hoorn, RAL. 2008. Papain-like cysteine proteases: key players at molecular battlefields employed by both plants and their invaders. *Molecular Plant Pathology*, **9**(1), 119–125.
- Sijwali, PS, Shenai, BR, Gut, J, Singh, A, & Rosenthal, PJ. 2001. Expression and characterization of the *Plasmodium falciparum* haemoglobinase falcipain-3. *The Biochemical Journal*, **360**(Pt 2), 481–9.
- Sim, YC, Lee, SG, Lee, DC, Kang, BY, Park, KM, Lee, JY, Kim, MS, Chang, IS, & Rhee, JS. 2000. Stabilization of papain and lysozyme for application to cosmetic products. *Biotechnology Letters*, **22**(2), 137–140.
- Soto-Mera, MT, López-Rico, MR, Filgueira, JF, Villamil, E, & Cidras, R. 2000. Occupational allergy to papain. *Allergy*, **55**(10), 983–984.
- Tastan Bishop, O, & Kroon, MC. 2011. Study of protein complexes via homology modeling, applied to cysteine proteases and their protein inhibitors. *Journal of Molecular Modeling*, **17**(12), 3163–72.
- Tillack, M, Nowak, N, Lotter, H, Bracha, R, Mirelman, D, Tannich, E, & Bruchhaus, I. 2006. Increased expression of the major cysteine proteinases by stable episomal transfection underlines the important role of EhCP5 for the pathogenicity of *Entamoeba histolytica*. *Molecular and Biochemical Parasitology*, **149**(1), 58–64.
- Tina, KG, Bhadra, R, & Srinivasan, N. 2007. PIC: Protein Interactions Calculator. *Nucleic Acids Research*, **35**(Web Server issue), W473–W476.
- Tovchigrechko, A, Wells, CA, & Vakser, IA. 2002. Docking of protein models. *Protein Science: a publication of the Protein Society*, **11**(8), 1888–96.
- Turk, B, Turk, D, & Salvesen, G S. 2002. Regulating Cysteine Protease Activity: Essential Role of Protease Inhibitors As Guardians and Regulators. *Current Pharmaceutical Design*, **8**(18), 1623–1637.
- Vajda, S, & Kozakov, D. 2009. Convergence and combination of methods in protein-protein docking. *Current Opinion in Structural Biology*, **19**(2), 164–70.
- Vakser, IA, & Kundrotas, P. 2010. Predicting 3D Structures of Protein-Protein Complexes. *Current Pharmaceutical Biotechnology*, 57–66.

- Vasiljeva, O, Reinheckel, T, Peters, C, Turk, D, Turk, V, & Turk, B. 2007. Emerging roles of cysteine cathepsins in disease and their potential as drug targets. *Current Pharmaceutical Design*, **13**(4), 387–403.
- Vernet, T, Tessier, DC, Chatellier, J, Plouffe, C, Lee, TS, Thomas, DY, Storer, AC, & Menard, R. 1995. Structural and Functional Roles of Asparagine 175 in the Cysteine Protease Papain. *Journal of Biological Chemistry*, **270**(28), 16645–16652.
- Wallner, B, & Elofsson, A. 2005. All are not equal: a benchmark of different homology modeling programs. *Protein Science: a publication of the Protein Society*, **14**(5), 1315–27.
- Wang, SX, Pandey, KC, Somoza, JR, Sijwali, PS, Kortemme, T, Brinen, LS, Fletterick, RJ, Rosenthal, PJ, & McKerrow, JH. 2006. Structural basis for unique mechanisms of folding and hemoglobin binding by a malarial protease. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(31), 11503–8.
- Wang, SX, Pandey, KC, Scharfstein, J, Whisstock, J, Huang, RK, Jacobelli, J, Fletterick, RJ, Rosenthal, PJ, Abrahamson, M, Brinen, LS, Rossi, A, Sali, A, & McKerrow, JH. 2007. The structure of chagasin in complex with a cysteine protease clarifies the binding mode and evolution of an inhibitor family. *Structure*, **15**(5), 535–43.
- WHO. 2009. *World malaria report 2009*. World Health Organization.
- Williamson, AL, Lecchi, P, Turk, BE, Choe, Y, Hotez, PJ, McKerrow, JH, Cantley, LC, Sajid, M, Craik, CS, & Loukas, A. 2004. A multi-enzyme cascade of hemoglobin proteolysis in the intestine of blood-feeding hookworms. *The Journal of Biological Chemistry*, **279**(34), 35950–7.
- Wuthrich, K. 1989. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science*, **243**(4887), 45.
- Xiang, Z. 2006. Advances in homology protein structure modeling. *Current Protein and Peptide Science*, **7**, 217–227.
- Zemla, A, Venclovas, Č, Moult, J, & Fidelis, K. 1999. Processing and analysis of CASP3 protein structure predictions. *Proteins: Structure, Function, and Bioinformatics*, **37**(S3), 22–29.
- Zhang, C, & Kim, SH. 2003. Overview of structural genomics: from structure to function. *Current Opinion in Chemical Biology*, **7**(1), 28–32.
- Zhang, Y, & Skolnick, J. 2005. The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(4), 1029–34.