In Silico Characterisation of the Four Canonical *Plasmodium falciparum* 70 kDa Heat Shock Proteins

A mini-thesis submitted in fulfilment of the requirement for the degree

of

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework / Thesis

in

Bioinformatics and Computational, Molecular Biology in the

Department of Biochemistry, Microbiology and Biotechnology

Faculty of Science

by

Rowan Hatherley

February 2012

ABSTRACT

The 70 kDa heat shock proteins expressed by *Plasmodium falciparum* (PfHsp70s) are believed to be essential to both the survival and virulence of the malaria parasite. A total of six Hsp70 genes have been identified in the genome of *P. falciparum*. However, only four of these encode canonical Hsp70s, which are believed to localise predominantly in the cytosol (PfHsp70-1 and PfHsp70-x), the endoplasmic reticulum (PfHsp70-2) and mitochondria (PfHsp70-3) of the parasite. These proteins bind and release peptide substrates in an ATP-dependent manner, with the aid of a J-domain protein cochaperone and a nucleotide exchange factor (NEF). The aim of this study was to identify the residues involved in the interaction of these PfHsp70s with their peptide substrates, their J-domain cochaperones and potential NEFs. These residues were then mapped to three-dimensional (3D) structures of the proteins, modelled in three different conformations; each representing a different stage in the ATPase cycle. Additionally, these proteins were compared to different types of Hsp70s from a variety of different organisms and sequence features found to be specific to each PfHsp70 were mapped to their 3D structures. Finally, a novel modelling method was suggested, in which the structures of templates were remodelled to improve their quality before they were used in the homology modelling process. Based on the analysis of residues involved in interactions with other proteins, it was revealed that each PfHsp70 displayed features that were specific to its cellular localisation and each type of Hsp70 was predicted to interact with a different set of NEFs. The study of conserved features in each PfHsp70 revealed that PfHsp70-x displayed various sequence features atypical of both *Plasmodium* cytosolic Hsp70s and cytosolic Hsp70s in general. Additionally, residues conserved specifically in Hsp70s of Apicomplexa, *Plasmodium* and *P. falciparum* were identified and mapped to the each PfHsp70 model. Although these residues were too numerous to reveal any information of specific value, these models may be useful for the purposes of aiding the design of drug compounds against each PfHsp70. Finally, the novel modelling approach did show some promise. Half of the models produced using the modified templates were of a higher quality than their counterparts modelled using the original templates. This approach does still require a lot of validation work and statistical evaluation. It is hoped that it could prove to be a useful approach to homology modelling when the only templates available are poor quality structures.

DECLARATION

I declare that this thesis is my own, unaided work. It is being submitted for the degree of Masters of Science in Rhodes University. It has not been submitted before for any degree or examination in any other university.

_____

_____

This _____ day of _____ 2012

DEDICATION

This thesis is dedicated to my sister, Holly, the source of my inspiration, and my motivation when times are tough.

- "Forever in my heart" -

ACKNOWLEDGEMENTS

I would like to acknowledge the following people for their contributions to this work:

My supervisor, Dr. Özlem Tastan Bishop, for her constant support and dedication to me as her student. She has always strived to ensure I stay motivated and encourages me to develop my knowledge and skills with respect to each of the various aspects of bioinformatics.

Mattys Kroon, for providing me with the original python scripts used in MODELLER and for showing me some of the ways in which they could be altered and improved.

My friends and former colleagues, Joyce Njoki Njuguna and Dustin Laming, for their support and friendship and always asking questions, which at times forced me to expand my own insight.

My friend, Crystal Clitheroe, for helping out with some of the proof-reading and providing support during the final stages of writing up.

My parents, Gavin and Susan, for their love and for always supporting me in every way they could.

ACKNOWLEDGEMENT OF FUNDING

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

I would also like acknowledge Rhodes University for providing funding through a Prestigious MSc Scholarship.

TABLE OF CONTENTS

CONTENTS OF SUPPLEMENTARY DATA

CHAPTER 2

Appendix A: Sequences and Alignments

Appendix B: MAFFT-PROMALS3D Comparisons

Appendix H: Interaction models (Directory)

File 1: Appendix H README.txt

File 2: PfHsp70-1.pse

File 3: PfHsp70-2.pse

File 4: PfHsp70-3.pse

File 5: PfHsp70-x.pse

Appendix I: By-Type models (Directory)

File 1: Appendix I README.txt

File 2: PfHsp70-2.pse

File 3: PfHsp70-3.pse

File 4: PfHsp70-x.pse

File 5: PfHsp70-1.pse

LIST OF FIGURES

LIST OF TABLES

LIST OF WEB SERVERS AND WEB-BASED APPLICATIONS

1. Genesilico online server

   https://genesilico.pl/toolkit/

2. HHpred online server

   http://toolkit.tuebingen.mpg.de/hhpred

3. HIV database consensus sequence maker

   http://www.hiv.lanl.gov/content/sequence/CONSENSUS/SimpCon.html

4. Mafft online server

   http://mafft.cbrc.jp/alignment/server/

5. NCBI BLAST

   http://blast.ncbi.nlm.nih.gov/

6. PlasmoDB

   http://plasmodb.org/plasmo/

7. PROMALS3D

   http://prodata.swmed.edu/promals3d/promals3d.php

8. ProSA

   https://prosa.services.came.sbg.ac.at/prosa.php

9. Protein interaction calculator

   http://pic.mbu.iisc.ernet.in/index.html

10. RCSB Protein Data Bank

    http://www.rcsb.org/pdb/home/home.do

LIST OF ABBREVIATIONS

110 kDa heat shock protein	Hsp110
15-Deoxyspergualin	DSG
40 kDa heat shock protein	Hsp40
70 kDa heat shock protein	Hsp70
90 kDa heat shock protein	Hsp90
Basic Local Alignment Search Tool	BLAST
Bcl2-associated athanogene	Bag
BiP associated protein	BAP
Discrete Optimized Protein Energy	DOPE
Electron microscopy	EM
Endoplasmic reticulum	ER
Endoplasmic reticulum Hsp70	ER-Hsp70
Endoplasmic reticulum associated degradation	ERAD
Escherichia coli	*E. coli*
Fast Fourier transform	FFT
Global distance test	GDT
Global distance test total scores	GDT_TS
Heat shock protein	HSP
Hidden Markov model	HMM
Immunoglobulin binding proteins	BiP
Malate dehydrogenase	MDH
Mitochondrial Hsp70 homologue	mtHsp70
Model quality assessment programs	MQAPs
Multiple sequence alignment	MSA
National Center for Biotechnology Information	NCBI
Nuclear Magnetic Resonance	NMR

Nuclear Overhauser effect	NOE
Nucleotide exchange factor	NEF
Parasitophorous vacuole	PV
Plasmodium falciparum	*P. falciparum*
Plasmodium falciparum 70 kDa heat shock protein	PfHsp70
Position-Specific Iterative BLAST	PSI-BLAST
Probability density function	PDF
Protein Data Bank	PDB
Protein interaction calculator	PIC
Red blood cell	RBC
Research Collaboratory for Structural Bioinformatics	RCSB
Root-mean-square deviation	RMSD
Saccharomyces cerevisiae	*S. cerevisiae*
Substrate binding domain	SBD
Three-dimensional	3D
Translocase of the inner membrane	TIM
Translocase of the outer mitochondria membrane	TOM
Weights sum of pairs	WSP

AMINO ACIDS

Listed below are amino acids with their respective triple letter codes (TLC) and single letter codes (SLC).

Amino Acid	TLC	SLC
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Iso	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

1. INTRODUCTION

1.1 MALARIA

Prevalence of malaria in sub-Saharan Africa results in the loss of thousands of lives every day (Chiang *et al*., 2009). This fatal disease is caused by parasites from the genus *Plasmodium* and persists in areas where climatic conditions are suitable for its insect host, the female *Anopheles* mosquito (Greenwood *et al*., 2008). Five species of *Plasmodium* parasites are capable of causing malaria in humans (Singh *et al*., 2004) and of these five, *Plasmodium falciparum* (*P. falciparum*) carries the highest mortality rate (Gardner *et al*., 2002). The phylum Apicomplexa, to which *P. falciparum* belongs, contains some of the most successful parasites of the present day (Striepen *et al*., 2007). Apicomplexan parasites cause diseases that are relevant to both human and livestock. There are many structural and mechanistic characteristics that are shared by members of this phylum. These include the presence of an apicoplast, the formation of a parasitophorous vacuole (PV) after host cell invasion, as well as ability to invade host cells and evade the immune system (Striepen *et al*., 2007). It is believed that the apicoplast is a plastid that is descended from a chloroplast, however contains many functions that are specific to the parasite (Waller & Mcfadden, 2004).

1.2 LIFE CYCLE OF *P. FALCIPARUM*

P. falciparum has a very complex life cycle, which involves the expression of various different proteins as it must survive in both a poikilothermic invertebrate, as well as an endothermic vertebrate host (Florens *et al*., 2002). The stage of the parasite's life cycle responsible for the pathology of malaria is the asexual stage in which the parasite divides and replicates within human erythrocytes (van Dooren *et al*., 2005). This stage begins with the parasite, as a sporozoite, entering the blood stream of its human host as the mosquito feeds (Hanssen *et al.,* 2010). These sporozoites enter host hepatocytes where they each develop into thousands of merozoites, which then invade the red blood cells (RBCs) of the human host. It is suggested that this is accomplished by means involving an actin-myosin motor system (Cowman & Crabb, 2002). Within the RBC, the parasite continues to develop through the ring, trophozoite and

schizont stages of its life cycle, feeding off haemoglobin as a protein source (Bannister *et al.,* 2000). Finally the RBC is burst, releasing 16 daughter merozoites, causing the fever associated with malaria (Hanssen *et al.,* 2010). Each daughter merozoite is capable of invading another RBC and repeating the cycle. This cycle takes place over a 48 hour time frame (Miller *et al.* 2002). The trophozoite stage is associated with some alteration of the cell's internal and external composition. In addition to the appearance of various parasitic membranous compartments, the parasite also modifies the RBC membrane in order to ensure adhesion to the vascular endothelium. This reduces their passage to the spleen where specialised macrophages would remove the infected cell (Hanssen *et al.,* 2010).

1.3 HEAT SHOCK PROTEINS

Molecular chaperones play key roles in the prevalence of various human diseases, including neurodegenerative disorders, cancer and diseases caused by intracellular parasites (Acharya *et al.* 2007). Heat shock proteins (HSPs) are most well-known for their ability to act as molecular chaperones, but may also serve other functions (Shonhai *et al.,* 2007). In the events involved in the life cycle of *P. falciparum,* mentioned above, HSPs play a variety of important roles. The expression of these proteins is largely linked to environmental stress (Kumar *et al.,* 1991). They are found in virtually all orders of life, conserved even across different kingdoms (Desai *et al*., 2010). A study by Subjeck *et al.* (1982) indicated that proteins that are most greatly up-regulated during times of stress are the 70 kDa, 90 kDa and 110 kDa HSPs (Hsp70, Hsp90 and Hsp110, respectively).

1.4 HSP70

The Hsp70 family is the most highly conserved of all the HSPs. Expression of these proteins is most often in response to heat shock and cellular stress in general (Kumar *et al.,* 1991). Hsp70s bind and release peptide substrates in an ATP-dependant manner, which involves interaction with proteins referred to as co-chaperones. In *Escherichia coli* (*E. coli*), the prokaryotic homologue of Hsp70, DnaK has been well-characterised in its role of folding nascent peptides (Hartl & Hayer-hartl, 2002). Two other paralogues to this protein, Hsc62 (Arifuzzaman *et al.,* 2002; Yoshimune *et al.,* 2002) and Hsc66 (Silberg *et al.,* 1998) have been identified. The respective roles of these proteins are the negative regulation of transcription and prevention of

protein aggregation. This gives an indication of the variable roles of Hsp70s, even at a prokaryotic level. In eukaryotes, an increased level of complexity is observed as Hsp70s are localised across separate organelles and display even more diverse functions (Desai *et al.,* 2010). The Hsp70s relevant to *P. falciparum* (PfHsp70s) are believed to localise primarily in the cytosol, nucleus, ER and the mitochondria of the parasite (Sargeant *et al.,* 2006).

1.4.1 TYPES OF HSP70S

1.4.1.1 CYTOSOLIC/NUCLEAR HSP70S

The primary function of cytosolic Hsp70s is to bind to and ensure the correct folding of nascent or denatured polypeptide chains (Hartl & Hayer-hartl, 2002; Kim *et al*., 1998). Cytosolic Hsp70s have been shown to interact with Hsp90 and Hop assist with targeting mitochondrial proteins to the translocase of the outer mitochondria membrane (Tom) complex in the initial stages of protein import into the mitochondria (Faou & Hoogenraad, 2011; Hoogenraad *et al.,* 2002). This cooperation between Hsp70 and Hsp90 forms part of many different cellular processes including the regulation of signalling pathways (Young *et al.,* 2001) and facilitates trafficking of proteins to the lysosome (Agarraberes & Dice, 2001). In yeast, cytosolic Hsp70s Ssa1 and Ssa2 have demonstrated a potential role in suppressing prion propagation (Jung *et al.,* 2000).

1.4.1.2 ENDOPLASMIC RETICULUM HSP70S

Hsp70s that localises in the endoplasmic reticulum (ER) were initially named a 78 kDa glucose regulated proteins, or Grp78, as these proteins are usually up-regulated in the absence of glucose (Pelham, 1986). In mammals, these proteins are termed immunoglobulin binding proteins (BiP) and in yeast are known as Kar2 proteins (Buck *et al*., 2007). For the purposes of this research, they will be collectively referred to as ER-Hsp70s. These proteins assist in all three main functions of ER chaperones (Buck *et al*., 2007). Primarily, ER-Hsp70s facilitate the translocation of proteins into the lumen of the ER (Vogel *et al*., 1990). Nascent peptides in the cytosol are bound either post-translation or whilst emerging from the ribosome, by ER-Hsp70s and are pulled into the ER lumen. It is believed that this is achieved through the action of a racheting mechanism (Matlack *et al*., 1999). Once the peptide has successfully entered the ER, the second function of these Hsp70s is to ensure correct folding and assembly. The final major function of

ER-Hsp70s is to target proteins for ER associated degradation (ERAD) by the cytosolic proteasome (Buck *et al*., 2007). This process is believed to be dependent on the action of both ER- and cytosolic Hsp70s. There has been some work done on the mechanistic aspects of ER-Hsp70 function. Blond-Elguindi *et al*. (1993) screened the affinity of a wide variety of peptides for ER-Hsp70, BiP, in order to explore the binding characteristics of the protein. They found that BiP preferentially bound peptides containing a motif comprised of alternating aromatic and hydrophobic residues. Another characteristic of ER-Hsp70s is the presence of an ER retention sequence (Munro & Pelham, 1987). This sequence was determined to consist of the residues KDEL in mammals, by Munro and Pelham (1987). Later, the motif HDEL was found to be the ER retention signal in yeast (Pelham, 1988). Anelli *et al*. (2002) also discovered that some ER proteins contain a terminal RDEL motif.

1.4.1.3 MITOCHONDRIAL HSP70S

The mitochondrial Hsp70 homologue (mtHsp70) has been referred to by a number of different names, including mortalin, PBP74 and GRP75 (Ran *et al.,* 2000). For the purposed of this chapter, however, the protein will be referred to as mtHsp70. The study of mtHsp70s has indicated that they are most closely related to the Hsp70s of proteobacteria and there are many motifs that are common only in these Hsp70s (Germot *et al.,* 1996). This is by no means surprising, as it agrees with the theory put forward by Mereschkovsky (1910) that suggests the origin of mitochondria to be the result of an endosymbiotic process, wherein a eukaryotic organism ingested a bacterial species (Andersson *et al.,* 2002). More recent research has indicated that aerobic mitochondria are all descendants of α-proteobacteria. It is thought that during coevolution, gene transfer events resulted in the majority of mitochondrial genes forming part of the genome of their eukaryotic hosts (Andersson *et al.,* 2002). The result of this is that almost all mitochondrial proteins are synthesised elsewhere and must be imported into the mitochondria (Bohnert *et al.,* 2007). One of the major functions of mtHsp70s is to form part of this import process (Azem *et al.,* 1997). The process mtHsp70 is involved in is specific to mitochondrial proteins that contain a 21 residue N-terminal presequence peptide required for protein import (Zhang *et al.,* 1999). Import of presequence peptides into the mitochondrial matrix is governed by the translocase of the inner membrane (TIM) complex (Bohnert *et al.,* 2007). This complex is made up of a number of different TIM proteins, as well at mtHsp70.

Studies involving the cellular localization of this mtHsp70s have shown interesting results. Although it was originally assumed that this protein was targeted purely to the mitochondria, Ran *et al*. (2000) have shown that mtHsp70s localise in both the cytoplasm and ER. The different cellular localizations of mtHsp70s are each associated with different functions, which are summarised by Wadhwa *et al.* (2002). These include a response to environmental stress, the facilitation of intracellular trafficking, antigen processing and control of the cell cycle, as well as the initiation of tumorigenesis (Wadhwa *et al.,* 2002).

1.4.2 STRUCTURAL FEATURES OF HSP70S

Borges and Ramos (2008) explain that all types of Hsp70s are divided into two major domains – the 45 kDa N-terminal ATPase domain and the 25 kDa C-terminal substrate binding domain (SBD). Additionally, it is stated that these domains are connected by a flexible linker region. Work done by this group on one of the human Hsp70 homologues (Hsp70.1), showed that the interaction of ATP with the ATPase domain caused conformational changes in the structure of Hsp70.1. Additionally it has been determined that the ATPase domain is involved in regulating the function of the SBD (Blamowska *et al*. 2010). Conversely, the binding of a protein substrate to the SBD stimulates the ATPase activity of the ATPase domain (Borges and Ramos, 2008). Research done by Bork *et al.* (1992) showed that the structure of the ATPase domain very closely resembles that of both the actin and hexokinase ATPase domains, indicating that the ATPase domain is highly conserved even across different protein families. An additional domain, the linker domain, which links the ATPase domain to the SBD of Hsp70, has been shown to be involved in the regulation of ATPase activity by binding to the IA and IIA subdomains of the ATPase domain (Swain *et al*. 2008). The 25 kDa SBD of Hsp70 has been further divided into two distinct sub-domains (Strub *et al*. 2003). The first is the β-domain, which contains the substrate binding pocket, where a protein substrate will be bound. The second sub-domain is the α-helical lid, located closest to the C-terminus of the protein. This α-helical lid is sub-divided into five different areas, denoted as A-E respectively, based on their relative positions from N-terminus to C-terminus (Strub *et al*. 2003, Worrall & Walkinshaw, 2007). The function of the α-helical lid is to regulate the functioning of the SBD. A graphical representation of these domains is presented in Figure 1.1, where it can be seen where these domains are situated within the protein.

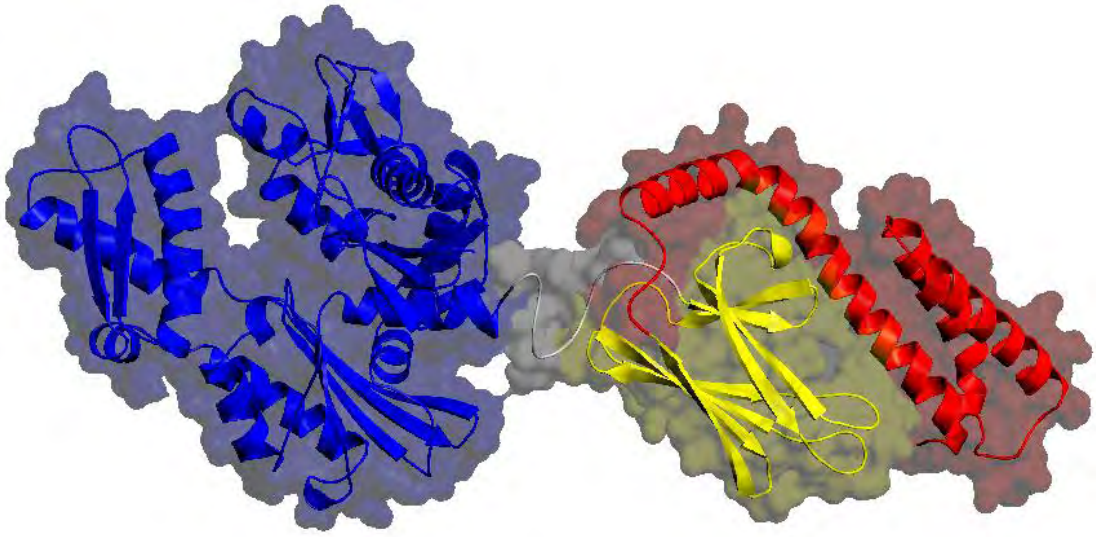

Figure 1.1: Structural features of Hsp70. This model represents the structure of one of the *P. falciparum* Hsp70s, PfHsp70-x. The protein was modelled using Bovine Hsp70 as a template using MODELLER. The model is coloured as follows. Blue: ATPase Domain; Grey: Linker Region; Yellow: β-subdomain of the SBD; Red: α-Helical Lid of the SBD.

1.4.3 Hsp70 ATPase Cycle

Hsp70s bind and release substrates in an ATP-dependent manner (Liu & Hendrickson, 2007). The functions of the ATPase domain and the SBD are inter-linked and the two domains are largely dependent on one another. When the ATPase domain is bound to ATP, the SBD loses affinity for its polypeptide substrate, as the α-helical lid opens, releasing the protein substrate (Liu & Hendrickson, 2007). With the α-helical lid in its open conformation, the SBD is able to bind to an unfolded protein substrate, where hydrophobic amino acid side chains are exposed (Hartl & Hayer-hartl, 2002). The event of substrate binding triggers the ATP hydrolysis at the ATPase domain, which causes the α-helical lid to close over the SBD, strongly binding the nascent polypeptide substrate (Hartl & Hayer-hartl, 2002; Mayer & Bukau, 2005). A nucleotide exchange factor (NEF) will then bind to the ATPase domain of Hsp70, which lowers the affinity of Hsp70 for ADP (Harrison *et al*., 1997). Due to the low nucleotide affinity of the NEF, ATP is transferred to the ATPase domain of Hsp70. In an ATP-bound state, Hsp70 has no affinity for its substrate and releases the bound peptide from its SBD (Dragovic *et al*., 2006).

1.4.4 J-DOMAIN PROTEINS

1.4.4.1 HSP40

Hsp40s are the eukaryote DnaJ homologue of approximately 40 kDa in mass. These are important co-chaperones of both cytosolic and ER-Hsp70s, as they are involved in stimulating ATPase activity and regulating substrate specificity of Hsp70 (Laufen *et al*., 1999). Interaction between this co-chaperone and Hsp70 is largely mediated through a conserved region of this protein known as the J-domain (Tsai & Douglas, 1996; Suh *et al.,* 1998; Jiang *et al.,* 2007). Additionally, PfJ4, a known *P. falciparum* Hsp40, which localises in the cytoplasm and nucleus of the parasite, has been shown to associate with PfHsp70-1 (Pesce *et al*., 2008).

1.4.4.2 PAM18

Pam18 is a J-domain protein found in the mitochondrial inner membrane and is associated with the ATPase activity of mtHsp70s (D'Silva *et al.,* 2003). Although this protein is only 18 kDa in mass, it appears to have a similar function to Hsp40 and it has been shown to be essential to the protein import function of mtHsp70.

1.4.5 NEFS

1.4.5.1 GRPE

GrpE is the NEF for the prokaryotic Hsp70 homologue, DnaK. This protein acts as a dimer to facilitate nucleotide exchange (Harrison *et al.,* 1997). Nucleotide exchange within mtHsp70s has been shown to involve a GrpE-like NEF, Mge1p (Azem *et al.,* 1997; Liu *et al.*, 2003). The interactions between GrpE and DnaK have been characterised by Harrison *et al.* (1997).

1.4.5.2 HSP110

Hsp110 have been described by Lee-Yoon *et al*. (1995) as the most highly diverged member of the Hsp70 family. The structure of Hsp110 closely resembles that of Hsp70; however Hsp110s differ most substantially to Hsp70 in their C-terminal domains, including the addition of a 100 residue acidic loop, not found in Hsp70s. Although Hsp110s have been reported as having

similar chaperone properties to Hsp70s (Oh *et al.*, 1997; Santos *et al*., 1998), differences in their functions have also been noted. Hsp110s have been shown to act as NEFs for cytosolic Hsp70s (Dragovic *et al*., 2006). Hsp110 is crucial to the ATP cycle required for the chaperone activity of Hsp70. The interactions between the yeast Hsp110 homologue, Sse1p, and Hsp70 have been characterised by Schuermann *et al.* (2008).

1.4.5.3 BAG DOMAIN PROTEINS

Proteins containing a Bcl2-associated athanogene (Bag) domain have been proposed as NEFs of cytosolic Hsp70s (Hohfeld & Jentsch, 1997). This is important, since no GrpE-like homologue has been found in the eukaryotic cytosol (Sonderman *et al.,* 2001). There is, however, evidence to suggest that Bag domain proteins cannot facilitate nucleotide exchange in Hsp70s (Bimston *et al*., 1998). The interactions between the Bag domain and Hsp70 have been characterised by Sonderman *et al.* (2001).

1.4.5.4 BAP

In terms of mammalian ER-Hsp70s, a BiP associated protein (BAP), has been identified as an NEF (Chung *et al*., 2002), however, the nature of its interaction with ER-Hsp70 has yet to be determined.

1.4.6 P. FALCIPARUM HSP70S

Within *P. falciparum*, six Hsp70 genes have been identified (Sargeant *et al*., 2006). These PfHsp70s have been named PfHsp70-1, PfHsp70-2, PfHsp70-3, PfHsp70-x, PfHsp70-y and PfHsp70-z (Shonhai *et al*. 2007). However, only four of these appear to be canonical Hsp70s (Pavithra *et al*., 2007; Shonhai *et al*., 2007). PfHsp70-y and PfHsp70-z are proposed to be Hsp110/Grp170 homologues. Although Hsp110s and Grp170s form part of the Hsp70 superfamily, they are a distinct sub-family of proteins whose functions differ from those of the canonical Hsp70s (Easton t *et al*., 2000). Of the PfHsp70s, only PfHsp70-1 and PfHsp70-2 have been studied in any detail. PfHsp70-1 is a cytoplasmic Hsp70 whereas PfHsp70-2 has been found to be expressed in the ER of *P. falciparum* (Kappes *et al*., 1993). Both these proteins are expressed in all blood stages of the parasite's lifecycle (Sharma, 1992). It is also reported by

Sharma (1992) that all of the other PfHsp70s are also expressed at some point in the blood stages of the parasite since patients' sera contained antibodies for these factors. Additionally, a recent study by Acharya *et al.* (2009) revealed that PfHsp70-x was detected in the clinical ring stage *P. falciparum.* This was an interesting finding, as the protein had never previously been detected in laboratory strains indicating that it may be involved in chaperone functions that are specific to interaction with the human host. The final canonical Hsp70 homologue, PfHsp70-3 is believed to localise in the mitochondria of the parasite, however this has yet to be confirmed (Sargeant *et al*., 2006).

1.4.6.1 PfHsp70-1 and PfHsp70-X

P. falciparum is believed to express two different cytosolic Hsp70s, PfHsp70-1 and PfHsp70-x (Sargeant *et al.,* 2006). PfHsp70-1 has also been found to localise in the nucleus (Kumar *et al* 1991). Of all the PfHsp70 proteins, PfHsp70-1 has received, by far the most research attention. It appears to be an inducible Hsp70 and is greatly up regulated in response to heat stress (Kumar *et al.,* 1991). *In vitro* studies of PfHsp70-1 reveal that the protein displays relatively high intrinsic ATPase activity compared to its human, bovine and bacterial counterparts (Matambo *et al.,* 2004). However, it was also discovered that the affinity of the protein for ATP was much lower than that of human Hsp70, suggesting that the activity of PfHsp70-1 is regulated by its NEF. Structural studies conducted by Shonhai *et al.* (2007) revealed that the mechanism of nucleotide exchange present in PfHsp70-1 may be similar to that of human heat-shock cognate 70 (Hsc70). Not only do these proteins contain identical residues at four of five contact points for two human NEFs, Bag-1 and Hsp70-binding protein 1 (Shomura *et al.,* 2002; Sondermann *et al.,* 2001), these residues are also predicted to be solvent exposed in PfHsp70-1 (Shonhai *et al.,* 2007). There are, however, no confirmed NEFs for any of the cytosolic PfHsp70s. PfHsp70-1 has shown the ability to reverse the thermosensitivity of an *E. coli* strain containing a partially defective DnaK (Shonhai *et al.,* 2005). It was unable to replace DnaK, however, as strains containing a truncated and non-functional DnaK were unable to grow at elevated temperatures. Interestingly, a chimeric protein composing of the ATPase domain of *E. coli* DnaK and the SBD of PfHsp70-1 was shown to be functional when expressed in a thermosensitive *E. coli* strain (Shonhai *et al.,* 2005). A similar effect was seen when PfHsp70-1 was expressed in yeast strains,

which expressed deficient Ssa1 and Ssa2 proteins (Bell *et al.*, 2011). Additionally, PfHsp70-1 assisted in other cellular processes, such as protein translocation and ERAD. Shonhai *et al.* (2005) also showed in their study the importance of the highly conserved linker region, between the ATPase domain and the SBD. Mutations in this region rendered the protein non-functional. *In vivo* experiments performed by Nicoll *et al*. (2007) revealed that the J-domains of two *P. falciparum* Hsp40s, PfJ1 and PfJ4, were able to interact with *E. coli* DnaK. There is evidence to suggest that PfJ4 is a co-chaperone of PfHsp70-1 (Pesce *et al.,* 2008). PfHsp70-1 has also displayed the ability to suppress thermally-induced malate dehydrogenase aggregation, reactivate denatured glucose-6-phosphate dehydrogenase (Misra & Ramachandran, 2009) Experiments involving the drug 15-Deoxyspergualin (DSG) suggested the possibility that PfHsp70-1 may be involved in trafficking proteins into the apicoplast (Ramya *et al.,* 2007). DSG is known to bind proteins containing an EEVD motif (Ramya *et al.,* 2006) and interferes with targeting proteins to the apicoplast. Apart from PfHsp70-1, only PfHsp90 contains an EEVD motif, which suggests either one or both of these proteins are involved in targeting nuclear-encoded proteins into the apicoplast (Ramya *et al.,* 2007). This supports work by Foth *et al*. (2003) that suggests protein transport into the apicoplast is facilitated by an Hsp70.

As indicated, PfHsp70-x is by far the less well-studied of the two cytosolic PfHsp70s. This protein was first detected by Kun and Muller-Hill (1989). A 201 residue peptide sequenced after it was discovered to react positively with sera from malaria immune patients, although was never identified as PfHsp70-x. The expression of the protein was only ever confirmed in 2009 by Acharya *et al.* (2009) when the proteome of clinical isolates of the parasite was determined using mass spectrometry. In the cytosol of higher eukaryotes is the traditional stress-induced Hsp70, as well as the constitutively expressed form of the protein, Hsc70 (Hartl & Hayer-hartl, 2002). Similarly, it is possible that PfHsp70-x is the constitutively expressed cytosolic Hsp70 of *P. falciparum* (Shonhai *et al*., 2007). Another interesting aspect of PfHsp70-x is the carboxy-terminal EEVN motif, since an EEVD motif is characteristic of cytosolic eukaryotic Hsp70s (Shonhai *et al*., 2007).

1.4.6.2 PfHsp70-2

PfHsp70-2 was the second Hsp70 to be discovered in *P. falciparum* (Kumar *et al.,* 1988; Peterson *et al.,* 1988). It was determined that the protein was expressed during all stages of the erythrocytic phase of the parasite's life cycle (Peterson *et al.,* 1988). A notable feature of PfHsp70-2 was the presence of a terminal SDEL motif, which suggested *P. falciparum* retention signals may be different to those found in mammals (Kumar *et al.,* 1988). Initial characterisation of PfHsp70-2was carried out by Kumar *et al.* (1991). It was determined that this protein localised in ER-like membranous structures found in the cytoplasm. However, unlike other ER-Hsp70s, PfHsp70-2 was not induced by partial glucose deprivation (Kumar *et al*., 1991).

1.4.6.3 PfHsp70-3

PfHsp70-3 has been shown to be phylogenetically related to other eukaryote mtHsp70s (Slapeta & Keithly, 2004), however, there has been almost no experimental work done on the protein. It has only been shown that that PfHsp70-3 may localise in the PV of the *P. falciparum* (Nyalwidhe & Lingelbach, 2006) and it has also been proposed that this protein is also targeted to the apicoplast (Foth *et al.,* 2003).

1.4.6.4 PfHsp70-y and PfHsp70-z as Hsp110s

PfHsp70-y and PfHsp70-z are two Hsp70 homologues of great interest. These two proteins have respective molecular masses of 100 kDa and 108 kDa (Shonhai *et al*. 2007). They both contain the highly conserved ATPase domain, but contain SBDs that differ greatly to the sequence conserved in other Hsp70s and the domains contain various inserts (Shonhai *et al*. 2007). These two also do not contain a nuclear localization signal as well as the linker region, two motifs that are conserved in the other PfHsp70s (Shonhai *et al*. 2007). Based on the size and architecture if PfHsp70-y and PfHsp70-z, it may be possible that these two proteins are *P. falciparum* Hsp110 homologues and based on their relative localizations (Sargeant *et al*., 2006), it is likely that PfHsp70-y and PfHsp70-z act as NEFs for PfHsp70-2 and PfHsp70-1/PfHsp70-x, respectively. Based on these observations, PfHsp70-3 would have no Hsp110 NEF, however, according to Harrison *et al*. (1997), this is likely, as it is explained that in eukaryotes, mtHsp70s are only associated with GrpE-like NEFs.

1.4.7 IMPORTANCE OF HSP70 TO THE *P. FALCIPARUM* LIFE CYCLE

It has been proposed that plasmodial Hsp70s are able to regulate actin polymerization (Tardieux *et al.,* 1998; Shonhai *et al*., 2007) and would thus be largely involved in the invasion of RBCs by the merozoite stage of the *P. falciparum* life cycle. When the parasite moves from its poikilothermic mosquito host into the endothermic human host, it experiences a temperature increase of more than 10^{o}C. The reason it is able to survive this severe temperature increase is that it expresses Hsp70 proteins, which maintain the integrity of all its other important proteins. The remodelling of the RBC by *P. falciparum* is thought to be largely regulated by Hsp70 proteins as these are involved in protein translocation across membranes (Gambill *et al*. 1993). Research by Banumathy *et al*. (2002) suggests that host Hsp70s may be recruited by the parasite for this purpose.

1.5 STRUCTURAL STUDIES

The function of a protein is thought to be related to its structure (Hegyi and Gerstein, 1999). It has been determined that knowledge of the structure of a protein enables the analysis of its function, interactions and antigenic behaviour (Krieger *et al.* 2003). Additionally the structure of a protein allows for the development of drugs that target that protein. Determining the three-dimensional (3D) structure of a protein by experimental procedures is often not possible, especially for larger proteins (Krieger *et al.* 2003). As a result, strategies have been developed in order to predict the 3D structure of proteins *in silico*. One such approach is known as homology modelling. It involves using the structure of a protein that has been solved experimentally as a template for predicting the structure of a protein whose structure is unknown (Xiang, 2006). If a suitable template is available, this is the most viable approach to take. Of the four canonical PfHsp70s, only PfHsp70-1 has a published structure based on homology modelling (Shonhai *et al.* 2008).

1.6 RESEARCH HYPOTHESIS

The *in silico* work done to date on PfHsp70-1 has involved the identification of residues important to substrate binding, as well as the interaction with potential J-domain proteins. Currently neither of these features has been studied in PfHsp70-2, PfHsp70-3 or PfHsp70-x. Additionally, no NEF has been identified for any of the PfHsp70 proteins and it has yet to be established whether or not PfHsp70-1, PfHsp70-2 or PfHsp70-x are likely to interact with an Hsp110 for this purpose. Additionally, there has been no structural work done on PfHsp70-2, PfHsp70-3 or PfHsp70-x. The current research aims to identify these sequence features as they occur in each PfHsp70. The sequences of these proteins will be analysed based on the different types of Hsp70s they represent and will be compared to similar types of Hsp70s expressed by different organisms. From this, sequence features will be identified that are conserved specifically in Apicomplexa and *Plasmodium* Hsp70s of each type. By modelling each of these proteins and mapping the specific sequence features to the models, it is hoped that these features can be studied at a structural level. Additionally, these proteins will be modelled in three different conformations, allowing these sequence features to be observed as they appear in different stages of each PfHsp70s ATPase cycle.

2. SEQUENCE ANALYSIS

Hsp70s expressed by *P. falciparum* are believed to localise in the cytosol, ER or mitochondria. These different types of Hsp70s are common to eukaryotes and have been found to develop functions specific to their cellular localisation. In this chapter, the four canonical *P. falciparum* Hsp70s, PfHsp70-1, PfHsp70-2, PfHsp70-3 and pfHsp70-x, were analysed based on their predicted cellular localisations. Sequence features relating to substrate binding and interactions with other proteins were identified in each PfHsp70 protein. Additionally, these proteins were compared with similar types of Hsp70s found in *Plasmodium*, Apicomplexa and eukaryotes, in order to identify regions that are conserved in *P. falciparum* and specific to Hsp70 from other parasitic organisms of the phylum Apicomplexa. Apart from PfHsp70-1 these residues had not previously been identified in the PfHsp70s. This is also the first time that putative NEF contact residues have been identified in any PfHsp70. It was found that each of the four PfHsp70s analysed had a number of contact residues that differed from the other PfHsp70s. It was also found that this trend was present when comparing different types of Hsp70 from different organisms. This indicated that the interaction of Hsp70s with peptide substrates and other proteins is, to some extent, specific to their cellular localization. When comparing the PfHsp70s to Hsp70s of the same type, found in other organisms, it was found that PfHsp70-x contains many sequence features that are atypical of *Plasmodium* cytosolic Hsp70s as well as cytosolic Hsp70s in general. It is therefore possible that PfHsp70 performs functions that are distinct from those of other cytosolic Hsp70s. Analysis of the different types of Hsp70s also revealed residues that were conserved specifically in the PfHsp70 proteins, as well as the *Plasmodium* and Apicomplexa sequences that were not present at a eukaryotic level. The various sequence features identified in this chapter are further studied at a structural level in the following chapter.

2.1 INTRODUCTION

2.1.1 SEQUENCE FEATURES OF INTEREST

In this chapter, residues of interest will primarily concern those involved with the interaction of each Hsp70 with their substrates as well as the interactions with other proteins. The SBD of

Hsp70 has been extensively studied and the residues which interact with the substrate were identified in *E. coli* DnaK, by Zhu *et al*. (1996). In addition, residues important in the formation of a hydrophobic arch and hydrophobic pocket during substrate binding have also been identified (Matthias P Mayer *et al*., 2000). As previously mentioned, the ATPase cycle of Hsp70 is regulated by two types of proteins. The first are J-domain proteins, which stimulates the ATPase activity of Hsp70. Residues thought to interact with the Hsp40 J-domain have been identified within the bovine Hsc70 sequence (Jiang *et al.,* 2005). Second are proteins involved in nucleotide exchange, which enable Hsp70 to release its substrate. The interaction between DnaK and GrpE has been well studied for many years (Brehmer *et al.,* 2001; Gelinas *et al.,* 2004; Harrison *et al.,* 1997; Żmijewski *et al.,* 2007) and has been used to further understand the interactions between mtHsp70s and Mge1 (Miao *et al.,* 1997; Sakuragi *et al.,* 1999). Although there appear to be multiple residues of DnaK that interact with GrpE, there have been six local areas of contact identified between these two proteins (Harrison *et al.,*1997). An important feature that has been determined is a loop within DnaK that is responsible for the interaction to occur. The key residue found within this loop corresponds to Gly32 of *E. coli* DnaK and Gly60 of *Saccharomyces cerevisiae (S. cerevisiae*) Ssc1. Mutation of this residue in DnaK (G32A and G32D) resulted in loss of binding between DnaK and GrpE, whereas in Ssc1, the mutation, G60D, allowed binding to Mge1p to occur, however nucleotide exchange was abolished (Harrison *et al.,* 1997; Miao *et al.,* 1997; Sakuragi *et al.,* 1999).

2.1.2 THE PROTEIN INTERACTION CALCULATOR

Although Harrison *et al.* (1997) have determined the contact points between GrpE and DnaK, only one of the residues involved in this interaction are reported. The identities of the other residues involved in GrpE-DnaK binding are not readily accessible. As a result, these residues must be determined by studying the structure submitted by Harrison *et al.* (1997). This can be done using the protein interaction calculator (PIC) server (Tina *et al*., 2007). This tool is able to determine different types of interactions, based on defined standards for each type of protein-protein interaction it calculates. Types of interactions that can be determined include disulphide bonds, interactions between hydrophobic residues, ionic interactions, hydrogen bonds and aromatic interactions (Tina *et al*., 2007).

2.1.3 BASIC LOCAL ALIGNMENT SEARCH TOOL

The Basic local Alignment Search Tool (BLAST) was designed by Altschul *et al*. (1990). It has been made available by the National Center for Biotechnology Information (NCBI) as an online search tool. It can be used to search for sequences that are most similar to the query sequence submitted, based on a specific substitution matrix. This not only takes sequence similarity and identity into account, but also penalises poorly substituted residues and gaps in the alignment. This tool has been further developed to identify more distantly related sequences by creating a profile, based on initial search results and performing iterative searches (Altschul *et al*., 1997). This is known as Position-Specific Iterative BLAST (PSI-BLAST).

2.1.4 SEQUENCE ALIGNMENTS

The alignment of protein or nucleic acid sequences is a useful tool in the field of bioinformatics. Even when comparing only two sequences, the use of a multiple sequence alignment (MSA) is favoured to a simple pairwise alignment. Aligning multiple sequences allows one to compare them as well as identify sequence features conserved during evolution (Chenna *et al*., 2003). Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment).

2.1.4.1 THE CLUSTAL SERIES

The Clustal series contains most widely used programs for global MSA (Chenna *et al*., 2003). These programs use a progressive alignment strategy, wherein an MSA is created by comparing two sequences at a time. The order in which these alignments are performed is based on a guide tree, created using phylogenetic-based algorithms. There have been many improvements to Clustal over the years. Up-weighting of more divergent sequences has been enabled in order to reduce biases caused by near-identical sequences (Thompson *et al*., 1994). Also, different alignment matrices can be used, depending on how divergent different sequences are. Position-specific gap penalties were introduced to favour gaps in likely loop regions, rather than regions that interfere with secondary structure (Thompson *et al*., 1994). More recent versions of Clustal use an alignment algorithm which incorporates dynamic programming (Chenna *et al*., 2003; Myers & Miller, 1988). They also use the neighbour-joining method (Saitou & Nei, 1987) to

generate the guide tree, which is slower, but far less error-prone than the UPGMA method originally used (Larkin *et al*., 2007). An additional feature of the latest versions of Clustal is an iterative refinement option, which incorporations a weights sum of pairs (WSP) score. Individual sequences are removed from the alignment and realigned, after which the WSP score is recalculated. If this score is reduced then the new alignment is accepted (Larkin *et al*., 2007).

2.1.4.2 MAFFT

The MAFFT MSA program was developed in order to reduce computational time, while retaining an accurate alignment (Katoh *et al*., 2002). The alignment algorithm is optimised to align residues based on volume and polarity. It has been determined that evolution of proteins favours substitutions which retain similar physico-chemical properties of this nature (Miyata *et al*., 1979). A fast Fourier transform (FFT) is used to convert the amino acid sequence to a sequence of vectors, representing the volume and polarity of each residue. Homologous regions between two proteins in an alignment are identified based on these values. MAFFT uses an improved, simplified scoring system that enables an alignment to be performed up to 100 times faster than other widely used MSA programs. MAFFT has developed a series of different alignment strategies for specific types of data sets (Katoh & Toh, 2008). Applicable to proteins with global homology, is the G-INS-I function (Katoh *et al*., 2005). This option uses iterative refinement to improve the accuracy of the alignment. This is done using the WSA scoring method, as well as an importance value, based on a matrix created by analysing gap-free segments of sequences (Katoh *et al*., 2005).

2.1.4.3 MUSCLE

The MUSCLE algorithm also produces fast multiple sequence alignments, with accuracy comparable to that of MAFFT (Edgar, 2004). The initial stages of alignment are similar to those of MAFFT, making use of two sets of progressive alignments. The first is based on a guide tree constructed using unaligned sequences and the second uses the aligned sequences to create a more accurate guide tree for progressive alignment. Refinement is then performed using a profile function (Edgar, 2004).

2.1.4.4 PROMALS3D

PROMALS3D is a program that takes both sequence and structure into account when aligning two sequences (Pei *et al*., 2008). This alignment tool creates a set of constraints, which it uses to guide the sequence alignment. The sequence-based constraints are created by creating a profile hidden Markov model (HMM), using alignment information from PSI-BLAST and secondary structure predictions from PSIPRED. The structural constraints are based on structural alignments of homologues with known structure, identified by PSI-BLAST or specified by the user. Using both sets of data has shown to greatly improve the accuracy of a sequence alignment (Pei *et al*., 2008).

The different sequence alignment programs discussed represent different options available when aligning sequences. The methods employed by these programs are each have their own merits and should be chosen to suit the type of alignment being performed. For the purposes of this chapter, the conserved nature of various sequence features is of interest and as such MAFFT presents itself as a useful alignment tool. This is because the algorithm MAFFT is designed to consider the physico-chemical properties of residues, which are conserved during evolution (Katoh *et al*., 2002; Miyata *et al*., 1979). In the following chapter, PROMALS3D becomes a more logical choice as it is designed to prepare sequence alignments for the purposes of protein modelling (Pei *et al*., 2008). It will be shown in this chapter, however, that due to the highly conserved nature of Hsp70s, there is little difference in aligning sequences with these two methods.

2.1.5 AIMS OF THIS CHAPTER

The aim of this chapter is to identify and compare sequence features which are important to each of the four canonical PfHsp70s. These features will initially be those described in section 2.1.5 above. Additionally, the PfHsp70s will be compared to other Hsp70s of their specific cellular localisation, as seen in eukaryotic sequences, apicomplexan sequences as *Plasmodium* sequences. This is done in order to reveal sequence features that are conserved specifically in each PfHsp70.

2.2 Materials and Methods

2.2.1 Sequence Acquisition

Amino acid sequences were obtained for *P. falciparum* (strain 3D7) Hsp70s from PlasmoDB (The plasmodium database collaborative, 2001). PfHsp70s sequences obtained included PfHsp70-1 (accession: PF08_0054), PfHsp70-2 (accession: PFI0875w), PfHsp70-3 (accession: PF11_0351) and PfHsp70-x (accession: MAL7P1.228). These were then divided into different types of Hsp70s, based on their presumed cellular localizations – cytosolic (PfHsp70-1 and PfHsp70-x), ER (PfHsp70-2) and mitochondrial (PfHsp70-3). For each type of Hsp70, three different groups were created. The first was a *Plasmodium* group. Representative proteins for each type were searched using protein BLAST. Cytosolic sequences were identified as having a terminal EEVD-like motif, whereas ER-Hsp70s were identified as having a terminal KDEL-like motif. For this purpose, PfHsp70-1 was used as the representative protein for cytosolic Hsp70s, as its localization has been experimentally confirmed. *P. falciparum* sequences were not included in the *Plasmodium* group. A similar search was done for apicomplexan Hsp70s. Again, *Plasmodium* sequences were not included in this group. Finally, a eukaryotic search was performed in a similar way. For the eukaryote search, a wide range of non-redundant sequences were searched. From the cytoplasmic search, each unique genus was searched to determine the type of organism it represented. The following representatives were obtained – human; rat; dinoflagellate; domycete; insect; amphibian; ichthyoid; mollusc; nematode; crustacean; plant; yeast and avain. When searching for ER-Hsp70s and mtHps70s, sequences of the same species were searched for specifically. If the species had no representatives, then the genus was searched, continuing by going up one order of classification at a time. Originally, prokaryotic sequences were searched for and grouped with the eukaryotic sequences, but these interfered with alignments. So instead all prokaryotic sequences were pooled together to form a separate group, consisting of nine different bacterial sequences and six archael sequences. A total of 90 non-PfHsp70 sequences were acquired for alignment. Refer to Table A.1 (Appendix A) for a list all sequences and accession numbers. The E-value of all sequences returned was 0.0, as reported by BLAST.

2.2.2 ALIGNMENTS AND CONSENSUS SEQUENCES

Alignments were performed using the MAFFT online server. Due to the highly conserved nature of Hsp70s, the G-INS-I alignment strategy was selected. Otherwise, default alignment options were used. For each localization group, alignments were performed for each specific classification group; e.g. cytosolic, eukaryotic Hsp70s. Each of these alignments was submitted to the HIV database consensus sequence maker. This was also done for the prokaryotic DnaK sequences. Finally an alignment was performed for each localization group, containing their specific PfHsp70, as well as the consensus sequences from their *Plasmodium*, Apicomplexa and eukaryote representatives, as well as the prokaryote DnaK group. These alignments were used to identify sequence features which are specific to each specific group. An alignment of the four canonical PfHsp70s was also performed. These sequences were annotated with sequence features found in literature searches, described in section 2.1.5, above.

Table 2.1: Lowest sequence identity pairs from each alignment. For each alignment performed, the sequence identity was calculated for each pair of sequences in the alignment. Listed below are the sequence pairs from each alignment with the lowest sequence identity, as well as the sequence identity for the pair.

Alignment	Sequence 1	Sequence 2	Sequence Identity (%)
All_DnaKs.fa	*R. bacterium* DnaK	*M. tarda* DnaK	51.62
Cytosolic_Alignment_Figure.fa	PfHsp70-x	Eukaryote_consensus	67.75
Cytosolic_All.fa	*A. thaliana* Hsp70B	*S. cerevisiae* Ssa1p	69.98
Cytosolic_Apicomplexa.fa	*T. annulata* Hsp70	*C. parvum* Hsp70	70.86
Cytosolic_Plasmodium.fa	*P. vivax* Hsp70	*P. yoelii yoelii* Hsp70	95.80
ER_alignment_Figure.fa	Eukaryotes_consensus	Plasmodium_consensus	64.12
ER_All.fa	*C. elegans* Hsp-3	*A. thaliana* BiP-2	64.55
ER_Apicomplexa.fa	*E. tenella* BiP	*B. rodhaini* Grp78	59.46
ER_Plasmodium.fa	*P. yoelii yoelii* ER-Hsp70	*P. vivax* Grp78	85.21
mt_Alignment_Figure.fa	Eukaryotes_consensus	PfHsp70-3	61.17
mt_All.fa	*E. tenella* mtHsp70	*A. echinatior* Hsc70-5	53.00
mt_Apicomplexa.fa	*T. annulata* mtHsp70	*T. gondii* mtHsp70	64.06
mt_Plasmodium.fa	*P. knowlesi* mtHsp70	*P. vivax* mtHsp70	97.44

An indication of the level of conservation between the sequences aligned is given in Table 2.1. The script **multi_seq_id_calculator.py** (Appendix D) was executed, which compared each sequence in the alignment with every other sequence in the alignment and calculated their identities. It then reported the lowest sequence identity pair for each alignment, with the sequence identity of the pair. Even amongst the prokaryotic DnaKs, the lowest sequence identity pair contained identical matching residues across at least 50% of the alignment. This high level of sequence conservation made it less challenging to obtain an accurate alignment and MAFFT does perform comparatively well when aligning sequences that are similar (Katoh & Toh, 2008). In order to further validate the use of the MAFFT alignments in this chapter, an additional set of alignments was created including the different types of eukaryotic Hsp70s, aligned with their respective PfHsp70s. This was done using both MAFFT and PROMALS3D. All sequences were then pooled into a single alignment such that the two alignment programs could be compared (refer to Appendix B, Figures B.1 – B.3). The ER-Hsp70 and mtHsp70 alignments (Figures B.2 and B.3, respectively) needed to be hand adjusted, as the highly variable N-terminal presequence peptides were aligned differently by the two alignment programs, thus distorting the rest of the alignment. The sequences were adjusted such that the GIDL motifs were aligned. Each protein was compared to itself as aligned by the two programs. In all alignments, the majority of the residues were aligned identically by both programs. Exceptions occurred in highly variable regions, where non-similar residues had to be placed on either side of a gap and both programs seemed to place residues in equally valid locations. It was decided that MAFFT was a suitable alignment program choice as it performed with very similar alignments to PROMALS 3D and its algorithm is well suited to the work of identifying conserved residues across different orders of life.

2.2.3 Hsp70 NEFs

Three different crystal structures of Hsp70s interacting with an NEF were obtained from PDB. First was 1DKG (Harrison *et al*., 1997) – This is the structure of *E. coli* GrpE, bound to the ATPase domain of DnaK. Second was 1HX1 (Sondermann *et al.,* 2001) – This is the structure of a human Bag domain in complex with bovine Hsc70 ATPase domain. Finally was structure 3C7N (Schuermann *et al.,* 2008) – The structure of the same bovine Hsc70 being bound by the yeast Hsp110, Sse1p. These three structures were each submitted to the PIC server to determine

likely NEF contact residues of each Hsp70. The Hsp70 sequences were retrieved from the PDB website and aligned to the four canonical PfHsp70 sequences. Residues identified by the PIC server were highlighted in each sequence where they occurred. The highlighted residues from each PfHsp70 were then highlighted in an identical set of alignments as those done for each Hsp70 type, described in section 2.1.1, above. These alignments were used to check whether or not the NEF binding sites were conserved in other Hsp70s. The Hsp110 and Bag domain binding sites were also checked against the cytosolic eukaryotic Hsp70 alignment (Figure A.1), whereas the GrpE binding sites were checked against the prokaryotic DnaK alignment (Figure A.10). Only highly conserved residues were considered as potential NEF binding sites.

2.3 RESULTS

2.3.1 ALIGNMENT OF PFHSP70S

The four canonical PfHsp70 sequences were aligned and annotated, based on sequence features found in literature. Sequence features considered included residues involved in substrate binding, as well as putative contact points for interaction with the J domain of an Hsp40 co-chaperone. Finally, sequence features known to be specific to each Hsp70 type, including the N-terminal transit peptide sequence for ER-Hsp70s and mtHsp70s, as well as the C-terminal tetrapeptides of cytosolic and ER-Hsp70s. The alignment of these sequences is presented in Figure 2.1. There are many residues that are common in all sequences, as shown by the high number identical and similar residues shown in black and grey, respectively. The proteins vary most in both their N- and C-terminal regions.

2.3.1.1 PUTATIVE J-DOMAIN CONTACT RESIDUES

Most of the residues of the putative J-domain protein contact residues, highlighted in blue, are similar. However, PfHsp70-1 contains a histidine residue at position 198, where the other proteins contain acidic residues. In Figure 2.3 it can be seen that this residue was conserved as an aspartic acid residue in all eukaryotic Hsp70 sequences collected. Also the contact point, corresponding to S398 of PfHsp70-1 is missing in PfHsp70-2 and PfHsp70-3. By referring to

Figures 2.3, 2.4 and 2.5, respectively, it can be seen that this point is highly conserved as a serine in cytosolic Hsp70s, but does not feature in any of the ER- or mtHsp70 sequences.

2.3.1.2 SUBSTRATE BINDING RESIDUES

Within the SBD, the hydrophobic arch residues (highlighted in yellow), are not conserved in all PfHsp70s. The first differs in each type of Hsp70, whereas the second only differs in PfHsp70-3. Again, by referring to alignments of each Hsp70 type, these substitutions are consistent for each Hsp70 type, with exception to the second hydrophobic arch residue of PfHsp70-2, residue Y450. Although this residue appears at this position in all apicomplexan ER-Hsp70s, alanine is conserved in the eukaryotic equivalents. Referral to the eukaryotic ER-Hsp70 alignment (Figure A.4) indicated that 11 of the 13 eukaryotic ER-Hsp70s contained alanine in this position. Tyrosine was found at this position in the other two. The substrate contact residue (highlighted in red) corresponding to position 450 of PfHsp70-1 also differs with each type of Hsp70. Once again the only PfHsp70-2 differs in this position to other Hsp70s of this type. The alanine is conserved in Apicomplexa, but this residue is highly conserved as threonine in the eukaryotic ER-Hsp70s.

2.3.1.3 C-TERMINAL TETRAPEPTIDES

Finally, a look at the terminal tetrapeptides of each of the PfHsp70s (boxed in a dashed line), as compared with the Hsp70 by type alignments, revealed that PfHsp70-x is the only cytosolic Hsp70 to have a terminal residue that is not aspartic acid. Also the 'SDEL' of PfHsp70-2 appears to be specific to *Plasmodium*, as no other ER-Hsp70 had this terminal repeat (refer to Figures A.4 and A.5).

2.3.2 NEF CONTACT RESIDUES

Likely Hsp70-NEF contact residues were determined submitting structures of NEF-binding to the PIC server and examining the level of conservation of these residues among other Hsp70s (Appendix C). Residues that were considered to be potential NEF contact points were highlighted in each of the four canonical PfHsp70s, represented in Figure 2.2. Highlighted in yellow are the putative GrpE contact points. Since mtHsp70s are supposed DnaK derivatives,

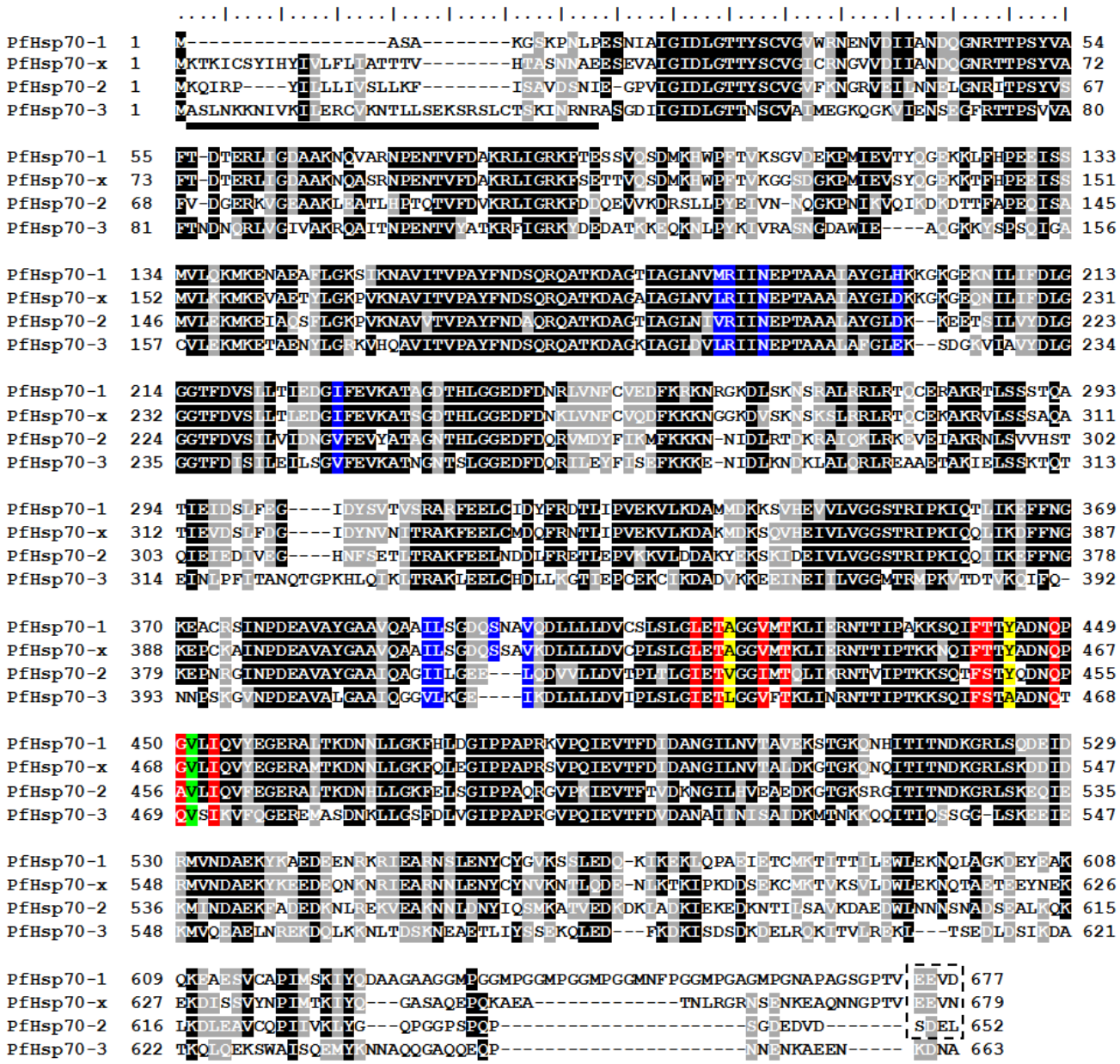

Figure 2.1: Alignment of the four canonical PfHsp70 sequences. The variable N-terminal transit motifs are underlined. Residues are highlighted as follows. Dark Blue: Putative Hsp40 contact residues; Red: Residues which interact with the substrate during peptide binding; Yellow and Green: Residues which form the hydrophobic arch and hydrophobic pocket, respectively; Black: Residues conserved at the level of identity; Grey: Residues conserved at the level of similarity. At the C-terminal end of the proteins, boxed in dashed lines are motifs characteristic of cytosolic and ER-Hsp70s

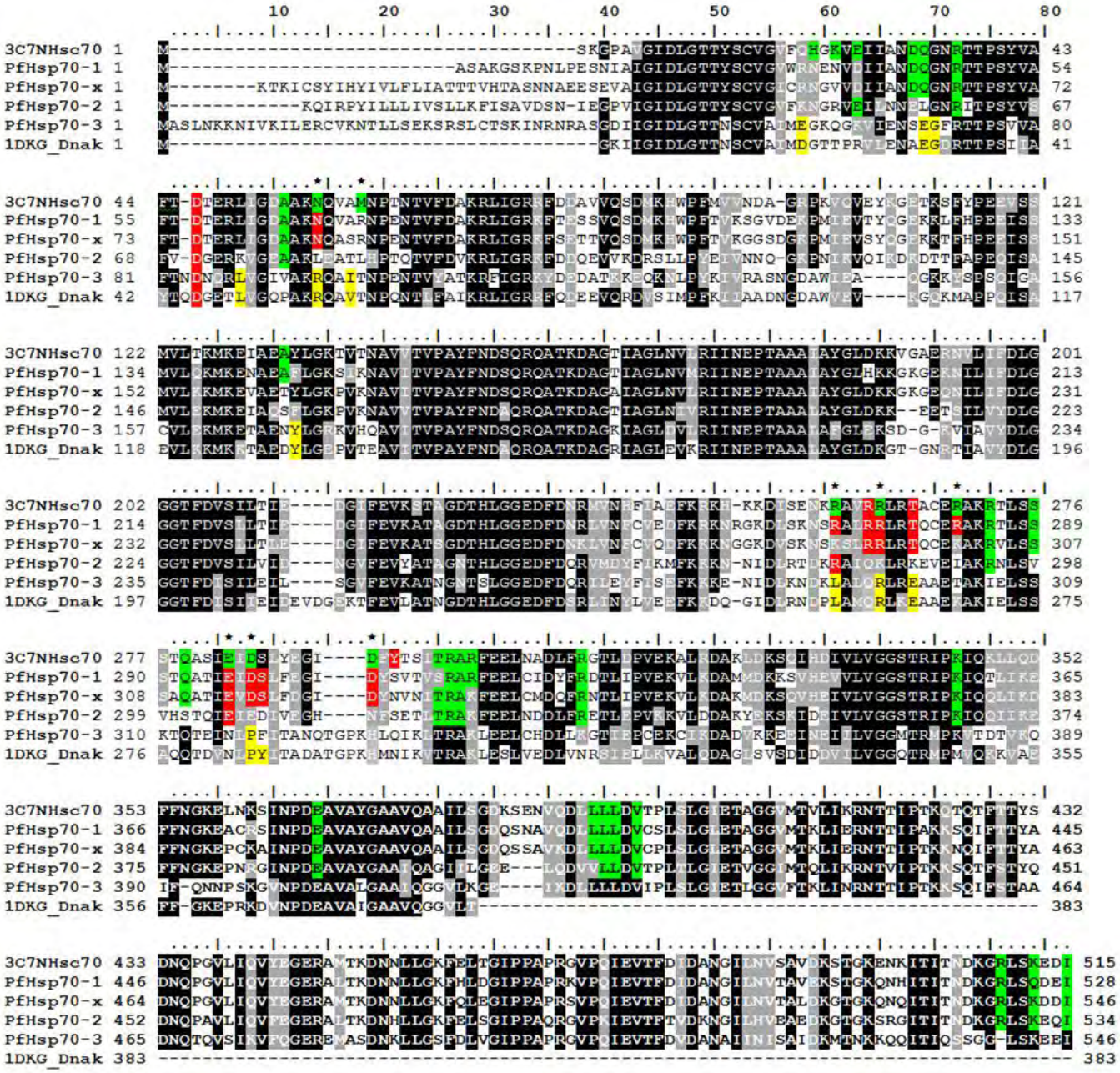

Figure 2.2: Putative NEF contact residues of each PfHsp70. The four canonical PfHsp70s were aligned to Hsp70s from three different crystal structures, each involved in interactions with a different type of NEF. Residues are highlighted as follows. Red: Putative Bag domain contact residues; Green: Putative Hsp110 contact residues; Yellow: Putative GrpE contact residues. Otherwise the residues are coloured as per Figure 2.1. Contact points common to both Bag and Hsp110 are denoted with an asterix.

these contact residues are most applicable to PfHsp70-3. As can be seen, PfHsp70-3 contains all except 3 of these residues. The exceptions are I95 and F319, which are conserved as valine and tyrosine, respectively, in eukaryotic mtHsp70s (Figure A.7). These are, however conserved substitutions. Residue E58 of PfHsp70-3, although different to the aspartic acid of bovine Hsc70, is conserved as a glutamic acid in eukaryote mtHsp70s. All but two of the Bag domain contact points are conserved completely in PfHsp70-1 and almost entirely in PfHsp70-x. The exceptions in PfHsp70-x are conserved substitutions of lysine in place or arginine at positions 289 and 300. Of the 33 positions considered as Hsp110 contact points, PfHsp70-1 contained 29 residues that were either identical or similar to those identified. PfHsp70-x contained 30 of such residues. PfHsp70-2 contains the least conserved NEF contact residues. For Bag domain and Hsp110 contact residues respectively, PfHsp70-2 contains 6 of 12 and 25 of 33 of these residues which are conserved at least at the level of similarity.

2.3.3 PfHsp70 Analysis by Type

For each type of Hsp70 (cytosolic, ER and mitochondrial), representatives were selected from different orders of life (eukaryote, Apicomplexa and *Plasmodium*). Consensus sequences were generated for each type of Hsp70 in each order of life and aligned to PfHsp70s of the respective type. A prokaryotic DnaK consensus sequence was added to this alignment.

2.3.3.1 Cytosolic Hsp70s

PfHsp70-1 and PfHsp70-x are considered to be cytosolic and so were aligned to the cytosolic Hsp70 consensus sequences. This alignment is shown in Figure 2.3, below. An initial notable feature in this alignment is the PfHsp70-x contains an N-terminal transit peptide region that is 18 residues longer than any other cytosolic Hsp70. Also notable is that the *Plasmodium* sequences contain a transit peptide which is highly conserved, except in PfHsp70-x. Residues highlighted in dark blue indicate residues that are unique to PfHsp70-x. Excluding transit peptide residues, PfHsp70-x contains 93 of these residues. PfHsp70-1 contains 14 of these residues (highlighted in red), 3 of which are shared by PfHsp70-x. PfHsp70-x also contains 32 residues that are conserved in eukaryotic, but not in other *Plasmodium* sequences (highlighted in yellow).

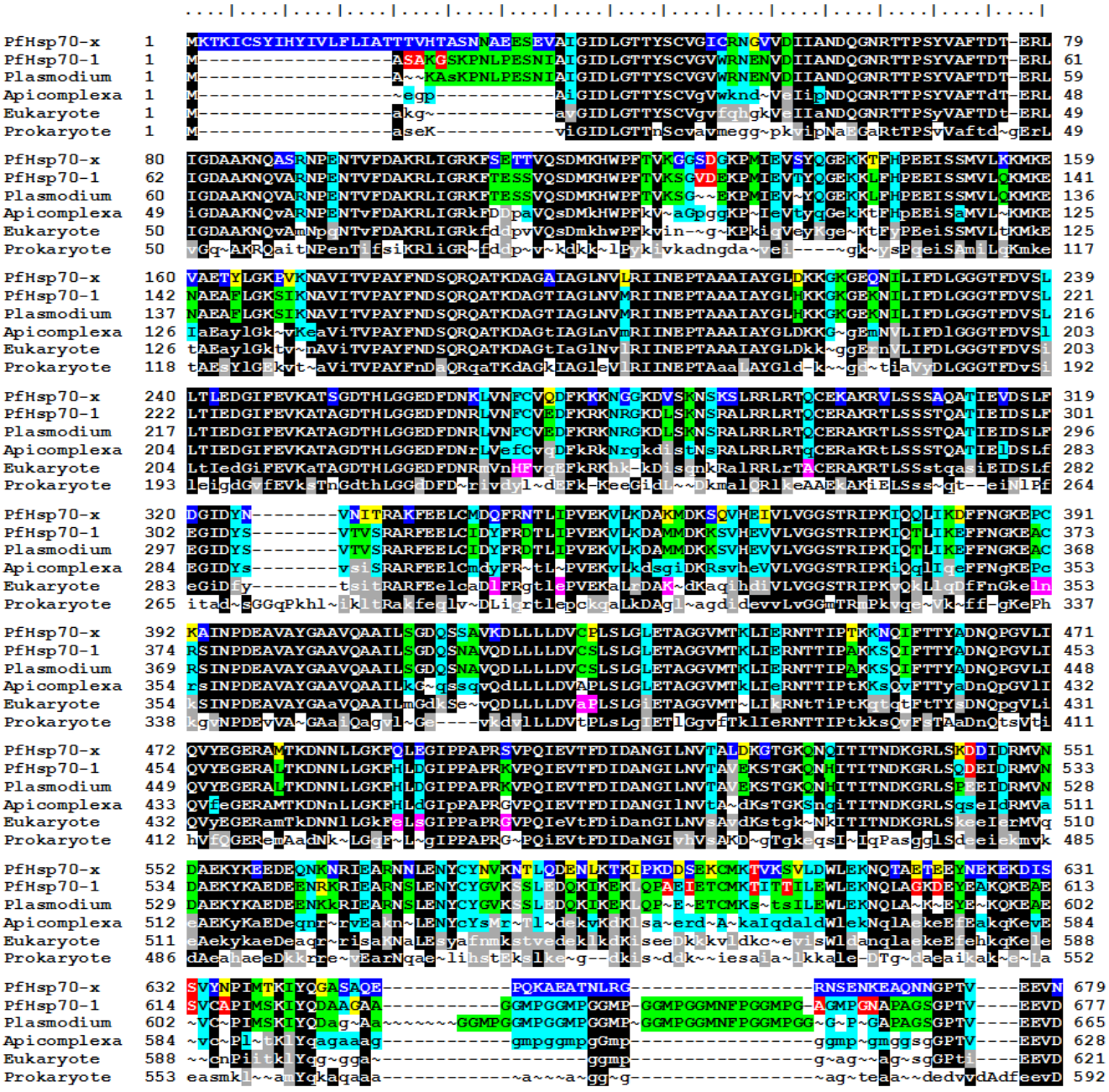

Figure 2.3: Cytosolic Hsp70 sequence alignment. PfHsp70-1 and PfHsp70-x were aligned with consensus sequences of cytosolic Hsp70 representatives of *Plasmodium*, Apicomplexa and eukaryotes, as well as prokaryotic DnaKs. Residues are highlighted as follows: Dark Blue: Unique to PfHsp70-x; Red: Unique to PfHsp70-1; Yellow: Residues present in either PfHsp70-x or PfHsp70-1 and eukaryotes, but not conserved in other *Plasmodium* Hsp70s; Green: Residues conserved in *Plasmodium*; Light Blue: Residues conserved in Apicomplexa; Purple: Residues which are highly conserved in eukaryotes, but not in PfHsp70-1 or PfHsp70-x. Consensus sequence letters in upper case were conserved in all sequences examined, whereas lower case letters indicate residues conserved in most sequences examined. A tilde (~) indicates a lack of consensus.

PfHsp70-1 contains 2 of such residues. There are many sets of residues which are specific to Apicomplexa (highlighted in light blue) and *Plasmodium* (highlighted in green), however often these are in variable regions with little or no conservation in eukaryote residues at these positions. Highlighted in purple are residues of interest. These residues are highly conserved in eukaryotes, but differ at their respective positions in either Apicomplexa or *Plasmodium* and hence PfHsp70-1 or PfHsp70-x. These residues were checked against the eukaryotic cytoplasmic Hsp70 alignment (Figure A.1). Conserved residues were considered at a level of similarity. Towards the C-terminus of the cytosolic Hsp70s is a region with two large gaps in the eukaryotic sequences. Analysis of the eukaryotic cytoplasmic Hsp70 alignment reveals this to be a glycine-rich region with at most one GGMP motif. In the apicomplexan Hsp70s, this is a repeated motif. This GGMP repeat is most prominent is Plasmodium sequences. In PfHsp70-x, however, this is neither a glycine-rich region, nor is there a single GGMP motif.

2.3.3.2 ER Hsp70s

PfHsp70-2 was aligned as one of the ER-Hsp70s. As can be seen in Figure 2.4, above, the N-terminal transit peptides are present in each sequence. This is a highly variable region, however some residues are conserved in *Plasmodium* and Apicomplexa. Apart from the transit peptide residues, PfHsp70-2 contains 26 residues that are not conserved in other proteins (highlighted in red). Almost all of these occur in variable regions where there is no conservation among the different ER-Hsp70 conservation sequences. PfHsp70-2 also contains 10 residues that are conserved in eukaryotic, but not in other *Plasmodium* sequences (highlighted in yellow). However, 8 of these residues occur where there are no conserved residues in the *Plasmodium* sequences. Again, residues of interest are highlighted in purple on the eukaryote conservation sequence. The residues, QN, highlighted in purple are not as highly conserved as other highlighted. However, at these two positions in the eukaryotic ER-Hsp70 alignments (Figure A.4), the residues are either QN, as in most sequences, or KG, as in the apicomplexan ER-Hsp70s.

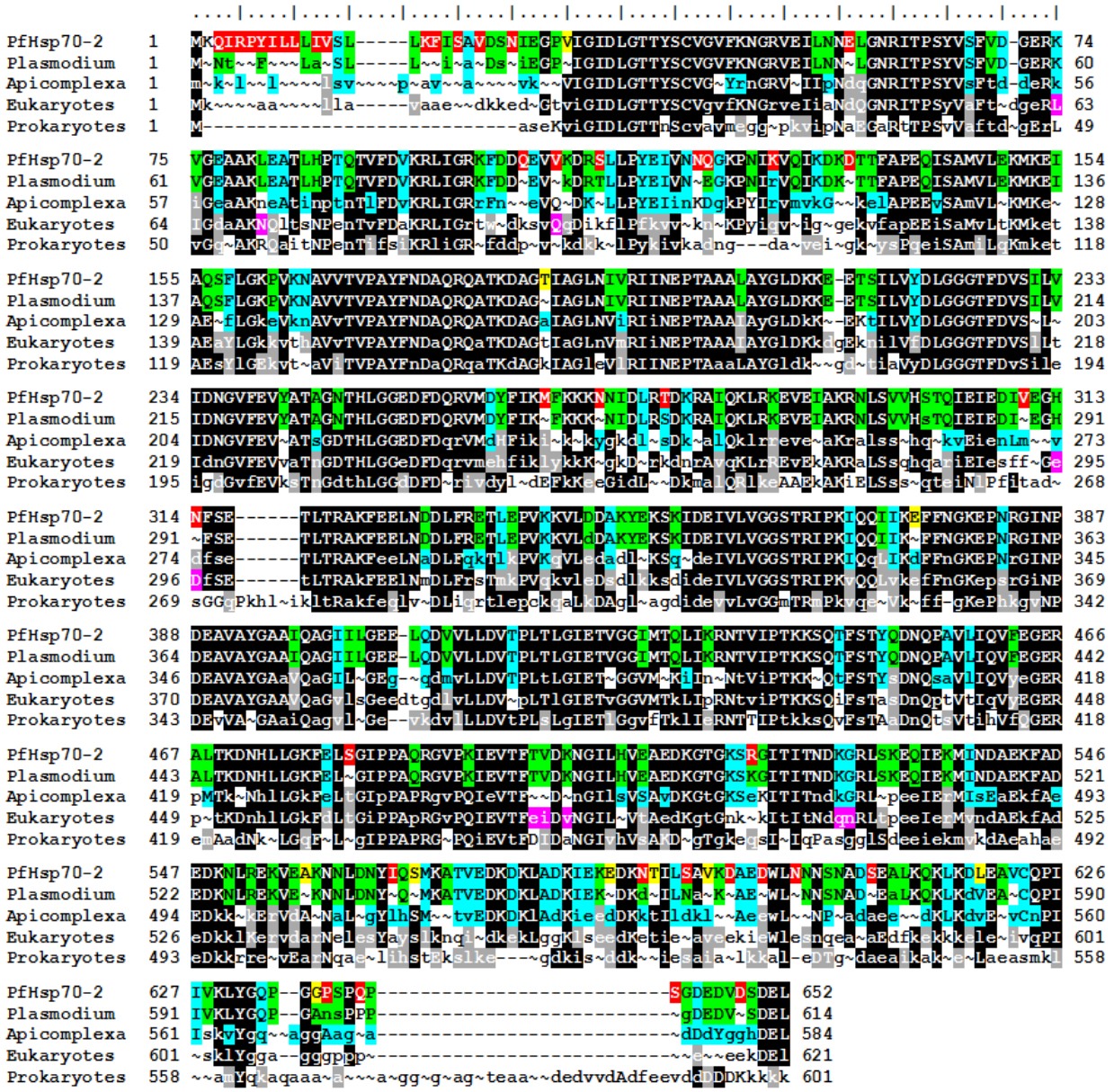

Figure 2.4: Alignment of ER-Hsp70 protein sequences. PfHsp70-2 was aligned with consensus sequences of ER-Hsp70 representatives from *Plasmodium*, Apicomplexa and eukaryotes, as well as prokaryotic DnaKs. Residues are highlighted as in Figure 2.3.

2.3.3.3 MTHSP70S

PfHsp70-3 was included in the alignment of the mtHsp70 sequences, shown in Figure 2.5. The transit peptide residues are highly conserved in the *Plasmodium* mtHsp70s. There were 16 residues that were specific to PfHsp70-3 (Highlighted in red) and 3 residues that were common

to eukaryotes and PfHsp70-3, but not other *Plasmodium* mtHsp70s. There was relatively little conservation across all eukaryotic mtHsp70s collected (Figure A.7), so no residues of interest could be identified. Within approximately the last 110 residues of the mtHsp70s, there are a large number of residues that are conserved, specifically in *Plasmodium* and Apicomplexa.

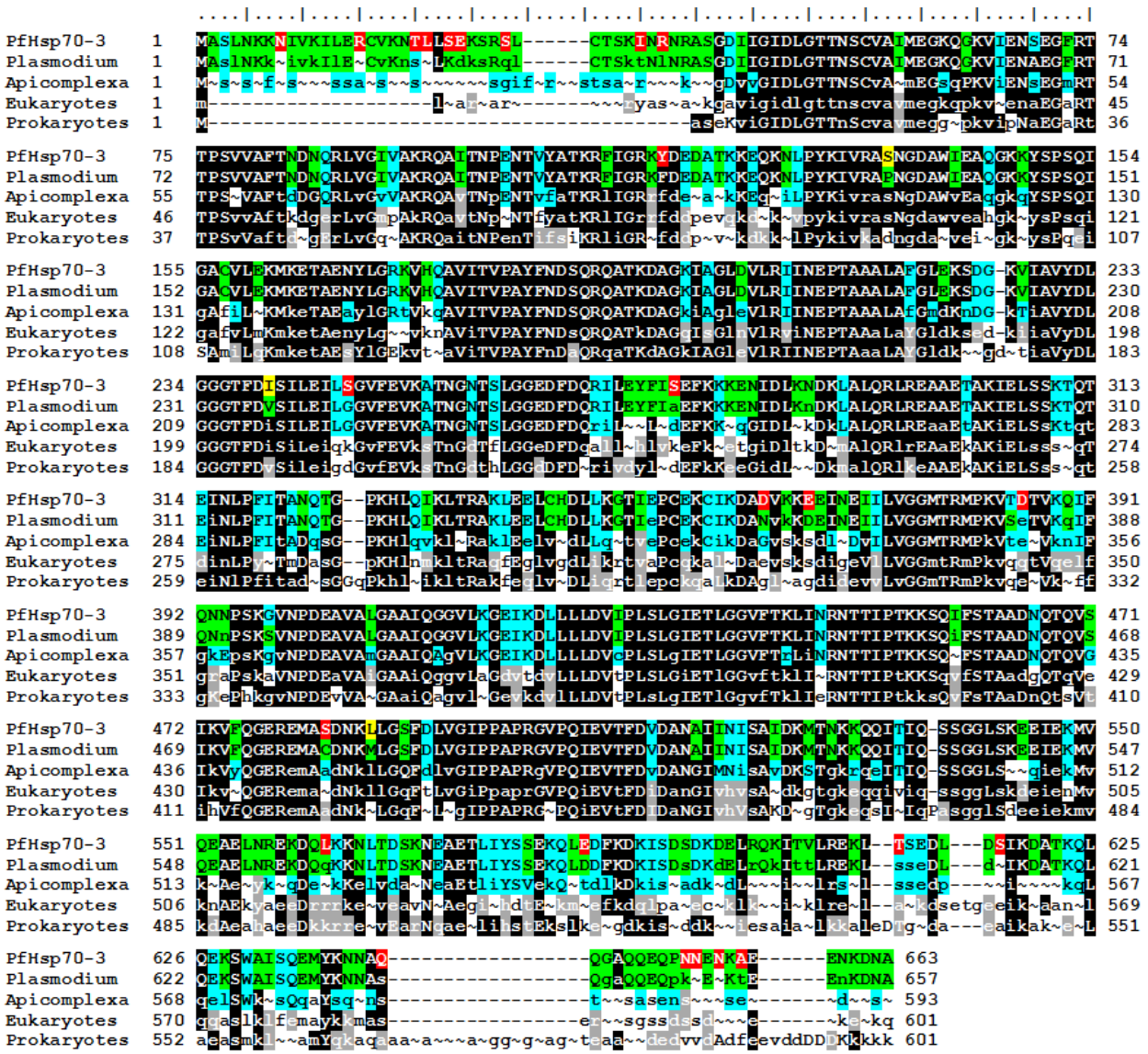

Figure 2.5: Mitochondrial Hsp70 sequence alignment. PfHsp70-3 was aligned with consensus sequences of mtHsp70 representatives of *Plasmodium*, Apicomplexa and eukaryotes, as well as prokaryotic DnaKs. Residues are highlighted as in Figure 2.3

2.4 DISCUSSION

2.4.1 ALIGNMENT OF PFHSP70S

The initial features analysed in the PfHsp70 sequences were residues involved in the interaction with the peptide substrate of Hsp70s, as well as contact points between Hsp70 and Hsp40 and NEFs.

2.4.1.1 HSP40 CONTACT RESIDUES

Putative contact points of interaction between *E. coli* DnaK and DnaJ have been previously identified (Suh *et al.,* 1998). As have those of bovine Hsc70 and the J-domain of its Hsp40 (Jiang *et al.,* 2007). In Figure 2.1, it can be seen that all of these contact points are present in PfHsp70-x, but two of them differ in PfHsp70-1 (M182 and H198), the latter of which is not a conserved substitution. By identifying these residues in Figure 2.3, it was discovered that both these substitutions are specific to *Plasmodium* sequences. Within the *P. falciparum* genome, at least 43 Hsp40s are coded for (Botha *et al.,* 2007). PfHsp70-1 has been shown to be capable of interacting with PfJ4 (Pesce *et al.,* 2008), as well as PF14_0359 (Botha *et al.,* 2011), both of which are cytosolic PfHsp40s. These substitutions indicate, however, that *Plasmodium* cytosolic Hsp70s interact with different J-domain residues to those of conventional cytosolic Hsp70s. Whether there are *P. falciparum* Hsp40s that are specific to PfHsp70-x or if PfHsp70-x interacts with host Hsp40s may be of interest. There is little information on the J-domain proteins that interact with ER-Hsp70s and mtHsp70s. Analysis of contact points reveals that PfHsp70-2 and PfHsp70-3 contain all residues conserved within their Hsp70 types, based on localisation. Both these proteins are missing the serine contact point, which reveals that the interaction of ER-Hsp70s and mtHsp70s with J-domain proteins may differ to that of cytosolic Hsp70s.

2.4.1.2 RESIDUES INVOLVED IN SUBSTRATE BINDING

Within the SBD, residues believed to be involved in substrate binding (Zhu *et al.*, 1996) as well as residues important to the formation of a hydrophobic arch and hydrophobic pocket during substrate binding (Zhu *et al.*, 1996; Mayer *et al.,* 2000) were highlighted within the PfHsp70 alignment. For each PfHsp70, these residues were compared with those found in Hsp70s form their type-specific alignment. Residues Y450 and A456 of PfHsp70-2 were found to be

conserved in apicomplexan ER-Hsp70s but rare in other eukaryotic ER-Hsp70s. These are hydrophobic arch and substrate contact residues, respectively. This implies that apicomplexan ER-Hsp70s may bind differently to peptide substrates than other eukaryote ER-Hsp70s do. Otherwise, it is possible that they bind a different range of peptide substrates. Across the different Hsp70-types, there are four substrate contact residues and two hydrophobic arch residues that differ. Rudiger *et al.* (1997) determined that the consensus sequence to which DnaK binds is four or five hydrophobic residues (most often leucine), flanked by basic residues. This differed from the motif found to be preferentially bound by the ER-Hsp70, BiP (Blond-Elguindi *et al.,* 1993), which contained large hydrophobic or aromatic residues ever second residue along the motif. Based on this, it is not unreasonable to assume that the sequence differences in the SBD residues of the Hsp70s are specific to the type of substrates bound in their cellular localization. It has been indicated by Zhu *et al* (1996), that the hydrophobic arch residues of DnaK, are often inverted in eukaryotic Hsp70s, but still form the required arch. This likely explains the positioning of these residues in PfHsp70-3, based on its presumed prokaryotic origins.

2.4.1.3 NEF CONTACT RESIDUES

Of the work done with regard to the interaction of the bacterial Hsp70 homologue, DnaK, with its NEF, GrpE, almost all literature relates back to work done by Harrison *et al.* (1997). While it is explained that there are six local areas of contact between these two proteins, the specific DnaK residues involved in this interaction are not revealed. As such, a study was performed to determine likely Hsp70 residues involved in this interaction. This was done by submitting crystal structures of Hsp70s interacting with different types of NEFs to the PIC server. This included the DnaK-GrpE crystal structure determined and studied by Harrison *et al.* (1997). The two other NEFs studied included a Bag domain (Sondermann *et al.,* 2001) and an Hsp110 (Schuermann *et al.,* 2008). When compared to Sondermann *et al.* (2001), this study identified an additional 2 residues that may potentially interact with the Bag domain. These residues correspond to R269 and D285 of Bovine Hsc70 from the crystal structure studied (1HX1). Both these residues were highly conserved in all eukaryotic cytosolic Hsp70s. Only residues M61 and Y294 were not conserved in PfHsp70-1, however, all apicomplexan cytosolic Hsp70s had identical substitutions at these points, with the exception of PfHsp70-x, which differed in the latter of these two. This

indicates that if apicomplexan cytosolic Hsp70s do interact with a Bag domain protein, these interactions differ in two of the conventional contact points. PfHsp70-2 contained only six of the Bag contact residues and it is unlikely that it is capable of interacting with a Bag domain protein. This also indicates that if PfHsp70-2 does interact with a BAP-like protein, as mammalian BiP does (Chung *et al.,* 2002), then this interaction is likely to take place across another distinct set of contact points. There were a total of 33 Hsp110 binding residues identified. Whether or not all of these are essential for interaction between Hsp70 and Hsp110 has yet to be determined. The number of contact points conserved in PfHsp70-1, PfHsp70-x and PfHsp70-2 does indicate that these proteins are likely to interact with PfHsp70-z and PfHsp70-y, respectively. The GrpE contact points were considered relevant to PfHsp70-3, since mtHsp70s interact with the GrpE-like NEF, Mge1p (Azem *et al.,* 1997). These points are all present in PfHsp70-3, with the exception of two conserved substitutions. It seems that the interaction between PfHsp70-3 and *P. falciparum* Mge1p is similar to that between DnaK and GrpE. An interesting observation is the degree of overlap between binding sites of the different types of NEFs. There are 12 sites where at least two types of NEF contact points overlap. It is likely that these are the residues that are most important to the nucleotide exchange process.

2.4.2 Analysis of Different Types of Hsp70s

2.4.2.1 Cytosolic Hsp70s

Analysis of the cytosolic Hsp70s revealed very interesting information about PfHsp70-x. This protein contains features that are atypical of both *P. falciparum* cytosolic Hsp70s and cytosolic Hsp70s in general. Firstly, PfHsp70-x contains an N-terminal transit peptide that is considerably longer than any other cytosolic Hsp70 considered. Renner and Waters (2007) explain that the purpose of a transit sequence is to dictate the localisation of the protein. Cytosolic Hsp70s typically do not have transit peptides. This suggests that PfHsp70-x does not necessarily remain in the cytoplasm and may be targeted elsewhere. Another feature is the large number of residues that are unique to PfHsp70-x. Many of these residues occur in highly conserved regions of the protein, where PfHsp70-x is the only sequence that differs amongst all cytosolic Hsp70s. Towards the C-terminus of the protein, all other cytosolic Hsp70s contained a glycine-rich region and all *Plasmodium* sequences contained at least five GGMP repeats, with the exception in both cases of PfHsp70-x, which didn't contain a single GGMP motif and only had two glycine

residues across this region of the protein. Finally, PfHsp70-x was the only cytosolic Hsp70 to have a terminal EEVN motif, whereas EEVD was conserved in every other cytosolic Hsp70. PfHsp70-1 on the other hand was very similar to all other *Plasmodium* cytosolic Hsp70s. Based on the alignment, there is 95.6% sequence identity between PfHsp70-1 and the *Plasmodium* consensus sequence, as calculated by BioEdit (Hall, 1999). An interesting feature of the *Plasmodium* cytosolic Hsp70s is the extended N-terminus when compared to apicomplexan and eukaryotic cytosolic Hsp70s. Although this is not as long as the extension seen in PfHsp70-x, it may be some form of a transit sequence. This region is conserved in the *Plasmodium* sequences, with the exception of PfHsp70-x.

2.4.2.2 ER-HSP70S

The ER-Hsp70s appear to be less conserved than the cytosolic Hsp70s and mtHsp70s. There are more areas where the *Plasmodium* and apicomplexan consensus sequences differ to the eukaryote consensus sequence. There are also more points where there is a different residue at the same position in each sequence. It is possible that ER-Hsp70s do not need to be as well conserved as other Hsp70 types. When the C-terminal tetrapeptide sequence of PfHsp70-2 was discovered, it was the first time an SDEL motif had been reported (Kumar *et al.,* 1988; Peterson *et al.,* 1988). Since this terminal tetrapeptide is an ER retention signal (Munro & Pelham, 1987), it was of interest that it differed to the highly conserved KDEL. This SDEL motif appears to be specific to *Plasmodium* ER-Hsp70s. It suggests that the mechanism by which sequences are retained in the ER is different in *Plasmodium* than other organisms.

2.4.2.3 MTHSP70S

The N-terminal transit peptide region is specifically conserved in *Plasmodium* mtHsp70s. Unfortunately, this region was not well represented in the eukaryotic sequences. Not all representative sequences were complete and some did not have transit peptides. The highly variable nature of this region resulted in the gaps of these sequences being more conserved than any given residue at most locations, when read by the consensus maker. As such it should be noted that the eukaryotic mtHsp70s did mostly have longer transit peptides than represented by the consensus sequence. Other than this, there are a large number of residues conserved specifically in *Plasmodium* and Apicomplexa in the final 110 residues of these sequences. This region may have functions specific to the *Plasmodium* at least.

3. HOMOLOGY MODELLING OF PFHSP70 PROTEINS

Determining the structure of a protein grants possible insight into how it functions and also allows for the development of drugs that target that protein. To this end, the four canonical PfHsp70s were studied at a structural level. All four proteins were modelled using three different templates, representing 3 different conformations the protein adopts. This enabled the study of the protein during different stages of the ATPase cycle. An attempt was made to improve the homology modelling process by improving the quality of the template structure by remodelling it. Although this approach showed some promise, more work needs to be done to optimise the process and determine the type of template structures it is best suited to, if at all. Once adequate models were obtained for each protein, the sequence features of each PfHsp70, identified in the previous chapter, were mapped to their structures. These models revealed that there is very little difference to the exposure of the J-domain and NEF contact as each PfHsp70 adopts different conformations. The conformation of the substrate binding residues does, however, appear to be dependent on the orientation of the alpha-helical lid, relative to the SBD. Further study of these sequence features may require the inclusion of structures of proteins that interact with PfHsp70.

3.1 INTRODUCTION

3.1.1 DETERMINATION OF PROTEIN STRUCTURE

One of the main purposes of structural biology has always been linking the structure of a protein to its specific function (Poulsen, 1994). The problem of determining the 3D structure of a protein has seen the development of a multitude of different approaches aimed at doing so. Each approach is classified on the basis of the information used to calculate the structure (Sali & Blundell, 1993). Primarily, approaches are divided into either experimental or hypothetical. Experimental approaches include X-ray crystallography, solution nuclear magnetic resonance (NMR) and electron microscopy (Foster *et al.*, 2007; Smyth & Martin, 2000; Topf & Sali, 2005). The theoretical methods are further sub-divided into physical and empirical approaches (Sali &

Blundell, 1993). Physical methods are based on satisfying thermodynamic constraints involved in protein folding (Gadja *et al*., 2010), whereas empirical methods predict 3D models based on other structures that have already been solved experimentally (Aszodi & Taylor, 1996). There are further sub-divisions of the empirical modelling approaches; however, comparative modelling produces the best results (Sali & Blundell, 1993).

3.1.2 EXPERIMENTAL METHODS OF PROTEIN STRUCTURE DETERMINATION

The RCSB Protein Data Bank (PDB) contains structural data regarding proteins whose structures have been solved experimentally (Berman *et al*., 2000). According to the PDB, no more than 8000 protein structures have been released in any given year, although the number of structures being generated is steadily growing every year.

3.1.2.1 X-RAY CRYSTALLOGRAPHY

The bulk of the protein structures solved experimentally are conducted using X-ray crystallography. Although this constitutes the most successful technique for determining protein 3D structure, the process is still very difficult (Larsen *et al.*, 1994). The process is only effective if a high yield of pure protein crystals is obtained. This can prove challenging, as the process of crystallization is complex and not entirely understood. Often, protein samples fail to produce crystals of suitable quality required (Smyth & Martin, 2000). During the process of X-ray crystallography, a crystalized protein sample is exposed to X-rays of a single wavelength that have been focused into a beam. This is done multiple times while the sample is rotated about a fixed axis and the way in which the X-rays are diffracted is recorded (Leslie, 2006). The diffraction data is also used to determine the dimensions and orientation of the unit cell, which is defined as the smallest repeating unit that makes up a crystal (Smyth & Martin, 2000). Each diffraction spot is also indexed and its intensity measured as it occurs in each diffraction image (Leslie, 2006). This data can then be used to calculate the amplitude and phase angle of the wave portion of the X-rays that form each diffraction spot (Smyth & Martin, 2000). Although the amplitude is simple to measure, the phase angle has to be determined indirectly. This is usually done by coating the protein with a heavy metal, whose amplitude and phase angles can be calculated. This is then used to determine the phase angles associated with the protein sample

(Smyth & Martin, 2000). Using the Fast Fourier transform (FFT), the amplitude and phase angle are converted to an electron density map. The electron density map is often refined to improve its quality make up for experimental error, after which the protein is modelled in a 3D conformation which best represents the data (Smyth & Martin, 2000). The model is also further refined based on energy minimisation and stereochemical constraints; all the while adhering closely to the electron density map. The refinements are tracked and recorded as an R-factor, which represents the difference between the experimental electron density data and that formed as a result of refinement (Smyth & Martin, 2000).

3.1.2.2 SOLUTION NMR

NMR-based approaches were once highly favoured, but the approach is limited to proteins which are less than 30 kDa in size (Poulsen, 1994). Because this technique is carried out in solution, one of its most appealing aspects is the ability to study the dynamic nature of proteins. This does, unfortunately, make it difficult to determine a high quality static structure, which is required for study. The use of NMR in protein structure determination takes advantage of the nuclear Overhauser effect (NOE) caused by nearby hydrogen atoms in the protein (Wüthrich, 2001). NOEs allow for the observation of hydrogen atoms, even in different residues, as long as they are within 5 Å of each other. The effect is specific to molecules that undergo Brownian motion in solution and is caused by the transfer of magnetism between two spins as they share a dipole-dipole moment (Solom, 1955). Another aspect of protein NMR is assigning peaks to specific amino acids (Wüthrich, 2001). Due to the relatively large size of protein molecule, a single NMR run will produce thousands of peaks for even small proteins. Additionally, there is a need to distinguish between peaks that are specific to the protein and those which are artifacts, caused by the technique itself (Moseley & Montelione, 1999). This is a very complex problem and although it can be solved manually by the use of Bayesian statistics, these days the data is interpreted computationally. An interesting new development is the use of so called in-cell NMR (Inomata *et al*., 2009). The use of cell penetrating peptides enables N^{15}-labelled proteins to be introduced into active cells, allowing protein NMR to be carried out *in vivo* (Inomata *et al*., 2009; Takayama *et al*., 2009).

3.1.2.3 ELECTRON MICROSCOPY

Techniques involving electron microscopy (EM) are relatively new with respect to determining protein structure. Originally, the main limitation was the resolution which was limited to approximately 30-50 Å (Nickell *et al*., 2006; Radermacher *et al*., 1992), as compared to X-ray diffraction, which is more than ten times higher in resolution. However, advances in this field are enabling better interpretation of EM data, allowing for the prediction of more accurate protein structures (Ahmed *et al*., 2011). Modern structures are being solved at a resolution of approximately 4.0 Å (Chen *et al*., 2011). EM is also extremely useful when visualizing cell architecture, rather than just individual protein molecules (Nickell *et al*., 2006). With respect to proteins, however, this technique is invaluable when visualising macromolecular protein complexes and protein multimers, such as virus capsids. An additional advantage of using EM is that, like solution NMR, the technique can be used to study proteins in their native environment (Nickell *et al*., 2006). This is achieved through a process of verification, wherein the prepared specimen enters a cryogenic state before the water molecules can enter a crystalline state (Aebi, 2003). Under the correct conditions, the water molecules remain in this state throughout the EM process. As such this technique is showing a great deal of promise.

3.1.3 THEORETICAL APPROACHES TO PROTEIN STRUCTURE DETERMINATION

The theoretical approaches to protein structure determination all revolve around the concept that the structure of a protein is determined by its amino acid sequence. When discussing these techniques, the protein whose structure is to be predicted by a specific method is referred to as the target. With respect to comparative modelling, a template refers to the structure of a protein that has been determined experimentally, which can be used to predict the structure of a target protein.

3.1.3.1 RMSD AND GDT TS

Before the theoretical approaches to protein modelling are discussed, it is worth defining two common methods of evaluating theoretical models. The first is measure of the root-mean-square deviation (RMSD) between a theoretical model and its experimentally determined counterpart. A mathematical definition of this value is given by Trott and Olson (2009). However, simply put, it

is a calculation of the average distance, in angstroms, between the positions of atoms of a theoretical model and their respective positions in the experimentally solved structure. The global distance test (GDT) is a measure of the percentage of a theoretical model that corresponds to its experimentally solved structure, within a specific cut-off distance (Zelma *et al*, 2001). The GDT total scores (GDT_TS) is the measure reported as the average of four separate GDT scores, calculated for the cut-off distances 1.0 Å, 2.0 Å, 4.0 Å and 8.0 Å, respectively (Zelma *et al*., 2001). These scores appear often in literature are used as a standard comparative score when benchmark tests are performed to compare different protein modelling approaches or model evaluation software.

3.1.3.2 *AB INITIO* MODELLING

The physical, theoretical methods of predicting the 3D structure of a protein are collectively referred to as *ab initio* methods. The ability to accurately predict the 3D structure of a protein, using nothing but the amino acid sequence is considered by many as the ultimate goal of structural bioinformatics (Klepeis & Floudas, 2003). A lot of work has gone into this particular area of research over the past few decades, but significant success has only ever been achieved with relatively small peptides. This method usually involves finding the global energy minima for the specific amino acid sequence being modelled (Bradley *et al*., 2005; Klepeis & Floudas, 2003). A major problem with this approach is that the computational requirements for modelling increases drastically as sequence length increases. Even a relatively small protein would take approximately 150 CPU days in order to create a single model predicting its structure (Bradley *et al*., 2005).

3.1.3.3 THREADING

Threading is partially an intermediate between *ab initio* techniques and homology modelling. This method predicts the folds of the query amino acid sequence and matches it up to a template structure which displays similar folding (Ngom, 2006). This method identifies templates for modelling, irrespective of the presence of homology or the level of sequence identify between the two proteins. The template sequence is used as a backbone structure in order to estimate the conformation of the protein of interest and *ab initio* modelling is used to determine the energy minima for the protein (Mirny & Shakhnovich, 1998). Although there has been some success for threading, this technique generally doesn't yield results that are as accurate as other methods.

3.1.3.4 HOMOLOGY MODELLING

Homology modelling predicts structures based on similar amino acid sequences. The structural features of a protein are conserved over even a large change to its sequence. A study by Illergård *et al*., (2009) took an in depth look at structural characteristics of proteins and found that a protein's structure is conserved up to ten times more readily than its sequence. As such, it is possible to accurately predict the 3D structure of a protein if there is sufficient similarity between the unknown protein of interest and a protein whose structure has been experimentally determined (Sali & Blundell, 1993). Although this is the most reliable form of *in silico* protein modelling, it still requires the use of other methods to make up for its limitations (Krieger *et al*., 2003). Sequence alignments need to be optimised such that they take structural data into consideration (Pei *et al*., 2008; Krieger *et al*., 2003). *Ab initio* techniques, as well as molecular dynamics are also required to best predict the structural features when the protein sequence of interest is not entirely covered by the template, as well as to ensure that the overall structure is in its most energetically favourable state (Krieger *et al.*, 2003; Sali & Blundell, 1993; Levitt, 1983).

3.1.4 PROCESS OF HOMOLOGY MODELLING

The process of homology modelling is probably best summarised by di Luccio and Koehl (2011), wherein the most problematic steps involved in this approach are discussed in depth. The main seven steps of homology modelling include; 1) Identification of a suitable template, 2) Alignment of the target protein sequence with that of the template, 3) Fitting the target protein sequence to the backbone of the template structure, 4) Modelling of loops, 5) Positioning of the side chains, 6) Refinement of the modelled structure and 7) Evaluation on the modelled structure. These steps are discussed in more detail below.

3.1.4.1 TEMPLATE IDENTIFICATION

As mentioned above, homology modelling relies on the use of a protein with a known structure and sequence that is sufficiently similar to the protein of interest. Rost (1999) describes a trend which can be used to infer homology between two proteins. According to this trend, the longer a protein sequence is, the lower the sequence identity required to confidently assume it can be used as a template for modelling a target protein. A more generalised figure is given by (Hillisch *et*

al., 2004), who claim that homology can be confidently assumed above 30% sequence identity and that below 15% sequence identity it can be confidently assumed that two sequences are not homologous. Further, there are studies that reveal more extreme examples where these rules do not apply. In 2008, (Roessler *et al*., 2008) identified two different Cro proteins, expressed by different strains of bacteriophage, which shared 40% sequence identity, yet exhibited different structural folds. Alternatively, Alexander *et al*. (2007) were able to design two different proteins, which shared 88% sequence identity, y*et al*so differed in both structure and function. Additionally, there have been many cases reported where proteins of low sequence identity adopt similar folds and display similar functions (di Luccio & Koehl, 2011).

These phenomena highlight the need to consider more than just sequence identity when searching for a suitable template. The HHpred server (Söding *et al.*, 2005) was designed as a highly sensitive search tool for finding protein homologues, even in cases of low sequence similarity. This can be used as an effective way of identifying suitable templates, largely due to its use HMMs. Instead of a substitution matrix, HHpred creates a profile HMM, based on MSAs performed by PSI-BLAST. This HMM contains information regarding the most likely residue at each position in the sequence, along with probabilities of the most likely substitutions, as well as probabilities associated with inserts or deletions occurring in a position specific manner. This is further improved by incorporating secondary structure predictions into the profile HMM. This fold-recognition aspect of the HMM ensures that homology is inferred from a structural perspective, in addition to taking the sequence identity into account. This increases the sensitivity of the search and increases the confidence that the template and target sequences are homologous (Söding *et al.*, 2005).

When evaluating a template, in terms of its suitability for use in homology modelling, there are three major aspects that must be taken into consideration. It is indicated by di Luccio & Koehl (2011) that these include i) the sequence identity between the template and target sequences, ii) the quality of the template structure (as given by its resolution in PDB) and iii) the query coverage. The query coverage simply indicates how much of the target sequence is accounted for within the 3D template structure. As long as these three variables are considered, a suitable template can be selected.

3.1.4.2 SEQUENCE ALIGNMENT

Venclovas (2003) indicated that many of the errors that arise during the homology modelling process are a direct result of incorrectly aligning the target protein sequence to that of the template. Correctly aligning two sequences for the purposes of comparative modelling can be more challenging as far less intuitive than standard sequence alignments. This is largely because an optimal alignment of the sequences of two proteins is not necessarily the best alignment for their structures (di Luccio & Koehl, 2011). As such, it is not enough to simply look at the sequences of the two proteins. The traditional sequence alignment technique needs to be further supplemented with a variety of other considerations. These include the prediction of secondary structure, structural alignments of known homologues, as well as information relating to the identification of conserved structural motifs, as well as residues which play crucial roles in the active sites of the proteins (di Luccio & Koehl, 2011).

3.1.4.3 FITTING TARGET SEQUENCE TO TEMPLATE BACKBONE

This step is where comparative modelling methods differ most and many different approaches have been established for doing so (Sali & Blundell, 1993). One of the more favoured methods was first proposed by Sali *et al*. (1990) and later incorporated into the homology modelling software, MODELLER (Sali & Blundell, 1993). The method is described as the satisfaction of spatial restraints, based on the sequence alignment between a template protein of known structure and a target protein to be modelled. The spatial restraints are presented in the form of a probability density function (PDF). The PDF is used to predict the most probable 3D structure of a protein, based on how its sequence aligns to that of the template. The relevant information used to construct these restraints was determined empirically by studying different sets of proteins of known structure from the same families. This includes structural properties of the protein that are governed by stereochemical constraints, such as Phi/Psi angles, Cα-Cα distances, etc., taking residue type into consideration. The protein is modelled in such that the molecular PDF is optimised (Sali & Blundell, 1993). It should be noted that MODELLER includes steps 4 and 5

(3.1.4.4 and 3.1.4.5, respectively) into its spatial restraints-based calculations. These sections, as described below, represent recent advances in these areas and not the methods used by MODELLER.

3.1.4.4 LOOP MODELLING

The flexible nature of loop regions of a protein makes it difficult to accurately predict their conformations (di Luccio & Koehl, 2011). These regions can however be essential to the correct functioning of a protein, however and so need to be modelled correctly. Currently, the methods that have shown the most accurate results involve the use of databases containing preferred conformations for a given loop sequence. This method assumes that throughout all proteins with the PDB there should be representative structural conformations of all short peptide sequences (Fidelis *et al*., 1994). This was shown to be mostly true for segments less than eight residues in length, but the chances of finding representative conformations for longer stretches became exponentially less likely as sequence length increased. Another problem identified with this approach was that at least three of the residues need to be aligned to the known structure in order to correctly orientate the additional loop structure (Fidelis *et al*., 1994). This approach has shown promise though as Baeten *et al*. (2008) demonstrated that if all sequence fragments of a protein are present in known structures, entire proteins could be reconstructed with a global RMSD of 0.48 Å, when compared to the native structure. An additional approach to loop modelling is to use *ab initio* methods, which are either derived empirically or involve minimising energy functions based on molecular mechanics (Olson *et al*., 2007). When these two methods of loop modelling are combined, loop regions of proteins can be accurately modelled up to nine residues in length (di Luccio & Koehl, 2011; van Vlijmen & Karplus, 1997).

3.1.4.5 POSITIONING OF SIDE CHAINS

The accurate positioning of the side chains of residues within a protein is both important and incredibly challenging (di Luccio & Koehl, 2011). This is most applicable to protein active sites or other sites where specific interactions take place, where the positioning of the side chains are crucial. It is due to the combination of all possible side chain conformations of each residue throughout a protein, which makes this calculation so challenging. To date, the best solution to

this problem involves the use of rotamer libraries (Xiang & Honig, 2001). The first rotamer library was built by Ponder and Richards (1987) and contained preferred side-chain conformations for each amino acid. The use of this library meant that fewer conformations had to be searched, making the process far less computationally expensive (Ponder & Richards, 1987). Many modern rotamer libraries have been developed and have been shown to produce accurate results when modelling side-chains (Xiang & Honig, 2001).

3.1.4.6 STRUCTURAL REFINEMENT

The current methods for refinement of protein structures are all based on an energy refinement approach proposed by Levitt & Lifson (1969). This involves refinement of the structure as a whole, using a generalised force field in order to bring the protein into its most stable conformation. Current approaches to structural refinement, when used for the purposes of modelling unknown proteins, have been shown to lower the quality of the models produced (di Luccio & Koehl, 2011). This is an area of modelling that is still being developed and there are currently conflicting opinions as to whether or not this step should be included in the modelling process (di Luccio & Koehl, 2011).

3.1.4.7 MODEL EVALUATION

As discussed previously, the determination of the 3D structure of a protein is far more reliable when using experimental techniques. As such, models produced by homology modelling need to be assessed in order to be validated. Model evaluation techniques are usually based on either physicochemical considerations or the comparison of theoretical models to native structures that have already been solved (Xiang, 2006). Unfortunately, there is no single method for adequately evaluating the structure of a modelled protein. Each has its own specific strengths and weaknesses and it is highly recommended that multiple methods are used for the purposes of model evaluation (Pawlowski *et al*., 2008). The model quality assessment programs (MQAPs) and related techniques that are relevant to this chapter are described below.

3.1.5 MODEL QUALITY ASSESSMENT PROGRAMS

3.1.5.1 VERIFY3D

Verify3D was developed by Bowie *et al*. (1991) after studying proteins of known structure. They developed a set of environments that they believed to be favoured differently by each amino acid. The environments were based on a set of three different characteristics – i) how buried the side chain on a residue is; ii) the fraction of the residue exposed to polar contact and iii) the local secondary structure where the residue resides. By combining these factors, 18 different environments were assigned. These environments were located in proteins of known structure and scores were assigned to each amino acid, based on the degree to which the environmental class was either favoured or disfavoured by that amino acid. When assessing a structure, Verify3D identifies the environmental class into which each residue in the model falls and scores it accordingly. The sum of these scores gives a reflection of the global quality of the model. The local quality of the model can be assessed by individual scores given to each residue. For graphical purposes, this score is often averaged over a window of 21 residues (Bowie *et al*., 1991; Luthy *et al*., 1992).

3.1.5.2 PROSA

ProSA, was developed by Sippl (1993) and is based on the Boltmann's principle (Sippl, 1990), which describes the forces that stabilise a protein in solution. From this, force fields are derived from known structures, based on protein-solvent interactions. Another aspect of these force fields is associated with the energy involved in the interaction of atoms of residues within the protein, expressed as a function of spatial separation (Sippl, 1990). The force fields are extracted from the structures in the form of mean force potentials, expressed as a function of amino acid sequence (Sippl, 1993). This enables force fields to be developed for models to be assessed. In order to establish the quality of the model, a polyprotein with an approximate length of 50 000 residues is constructed from short fragments of known protein structures, using only backbone atoms. Atomic distances and steric considerations are taken into account when constructing this polyprotein, in order to ensure that any fragment taken from it is in a reasonable conformation. The backbone of the of the model in question is hidden within the polyprotein and can only be identified as a fragment whose score corresponds to that of the original model, when assessed

using the technique mentioned above. A hide and seek method (Sippl, 1993) is used to search the polyprotein by shifting along the protein, assessing the mean force potential of fragments the same length of the model in question. These calculations are used to establish a Z-score, which describes how the mean force potential of the model deviates from that of other random conformations taken from the polyprotein (Sippl, 1993). In the current online version of ProSA (Wiederstein & Sippl, 2007), the Z-score for the model assessed is plotted along with those of those of all known structures in the PDB as a function of sequence length. If the model Z-score falls outside the range established by native structures it is considered to be inaccurate. Additionally, the mean force potential of the model is plotted as a function of its amino acid sequence, averaged across both 10 and 40 residue windows (Wiederstein & Sippl, 2007).

3.1.5.3 METAMQAPII

MetaMQAPII was designed with the goal of accurately assessing the local structural quality of a model in addition to global accuracy, as other MQAPs perform relatively poorly with respect to this aspect of model evaluation. This program is a meta-predictor designed by Pawlowski *et al*. (2008) that incorporates other MQAPs in order to increase the accuracy of the results. The eight MQAPs used by MetaMQAPII include Verify3D, ProSA, ANOLEA (Melo & Feytmans, 1998), BALA-SNAPP (Krishnamoorthy & Tropsha, 2003), TUNE (Lin *et al*., 2002), REFINER (Boniecki *et al*., 2003) and PROQRES (Wallner & Elofsson, 2006). Like Verify3D, MetaMQAPII assesses each residue in a structure by first placing in a group. However, MetaMQAPII will assign the residues to one of 315 groups, rather than the 18 environments used by Verify3D. The groups were established by assessing the structures submitted to the CASP 5 and CASP6 experiments, in the category of template-based modelling. These were approximately 8250 models, which predicted the structures of 84 target proteins. Each residue of each model was grouped as followed. First, the models were assessed using ProQ (Wallner & Elofsson, 2006) and given a global accuracy score. Based on these scores, each model was divided equally into one of seven bins – Bin 1 containing residues from the worst-scored models, while bin 7 contained residues of models with the best ProQ scores. Secondly, in each bin, residues were further divided equally into another one of five bins, based on the extent to which they were buried within their respective structures, as determined by ResDepth. After this, each residue was further subdivided into one of three bins, based on residue type – hydrophobic,

hydrophilic and other. Finally, each residue was then further subdivided into another one of three bins, based on local secondary structure. A combination of all these bins yielded 315 groups into which a residue of a structure could be divided. For each of these groups, a unique linear regression model was developed to determine the RMSD of a residue in that group from its location in the native structure, based on how it was scores by the combination of eight different MQAPs. Due to the use of a linear regression model to predict the outcome of a non-linear relationship, the models use ranking scores from 0 – 100, rather than an RMSD values in angstroms. These ranking scores can, however be converted back to an RMSD value. Upon completion of evaluating a model, MetaMQAPII returns a predicted GDT_TS score, which gives a prediction of the global quality of the model, as well as a predicted global RMSD value, describing the predicted deviation in angstroms from the true protein structure. Additionally, the residue scores of each MQAP used in the assessment is returned. Finally, a PDB coordinate file of the model is returned, wherein the B-factor for each residue is replaced with the ranking score from the meta-predictor calculation. This can be used to give a graphical representation of the model and easily identify problematic regions (Pawlowski *et al*., 2008).

3.1.5.4 DOPE SCORE

DOPE stands for Discrete Optimized Protein Energy (Shen & Sali, 2006). This score represents a statistical potential that was derived by studying proteins from the PDB. Based on PDFs calculated from native structures, the DOPE score is dependent on inter-atomic distances of all atom pairs within a structure. When compared to other scoring functions, DOPE score performed best at identifying the best model in a model set and displayed the most accurate correlation between score and RMSD (Shen & Sali, 2006). DOPE score was incorporated into MODELLER when version 8 was released. A normalised DOPE score (DOPE Z-score) has also been established, which takes the length of the structure into account, making the score comparable to that calculated for other protein models (Sali, 2009).

3.1.6 AIMS OF THIS CHAPTER

As indicated above, the process of obtaining a 3D model of a protein, regardless of the method used, can be a long and arduous process if done correctly. Homology modelling involves many steps, for each of which several approaches have been developed. In addition to this, the technique grants less certainty of an accurate model being produced when compared to experimental approaches. The overall goal of this chapter will be to produce high quality 3D models for each PfHsp70 protein, such that they are suitable for *in silico* functional studies. The modelling program used in this chapter will be MODELLER release 9v7 (Mod9v7). As such, the initial focus of this chapter will be targeted towards determining the effect that the use of this program will have on the modelling process. Specifically, whether or not errors are introduced based on the way it is designed to model proteins. Also considered, are the different modelling options offered by MODELLER and which are most effective. Additionally, a novel modelling approach is introduced, wherein the template structures are remodelled before they are used in the process of modelling the PfHsp70 proteins. This approach is compared to the traditional modelling approach of using template structures as they appear in the PDB. Between these two approaches, the top models are selected and further evaluated in order to validate their use in future studies of these PfHsp70 proteins. Additionally, this chapter will focus on sequence features identified in the previous chapter, as they appear within the protein at a structural level.

3.2 METHODOLOGY

The methodology implemented in the modelling process of this chapter is summarised in Figure 3.1 below. Templates used in modelling were identified. These were then remodelled, creating two groups of templates; 1) the original templates and 2) the remodelled or "modified" templates. After the relevant sequence alignments were performed, the proteins were modelled separately using both the original and modified templates. All models produced underwent loop refinement, in order to attempt to improve areas of the models which showed poor local quality, as identified by reviewing MetaMQAPII results of each model in PyMOL. All models produced were ranked according to their DOPE Z-scores. The top models were submitted to MetaMQAPII for further evaluation. All python scripts referred to in this chapter are present in Appendix E.

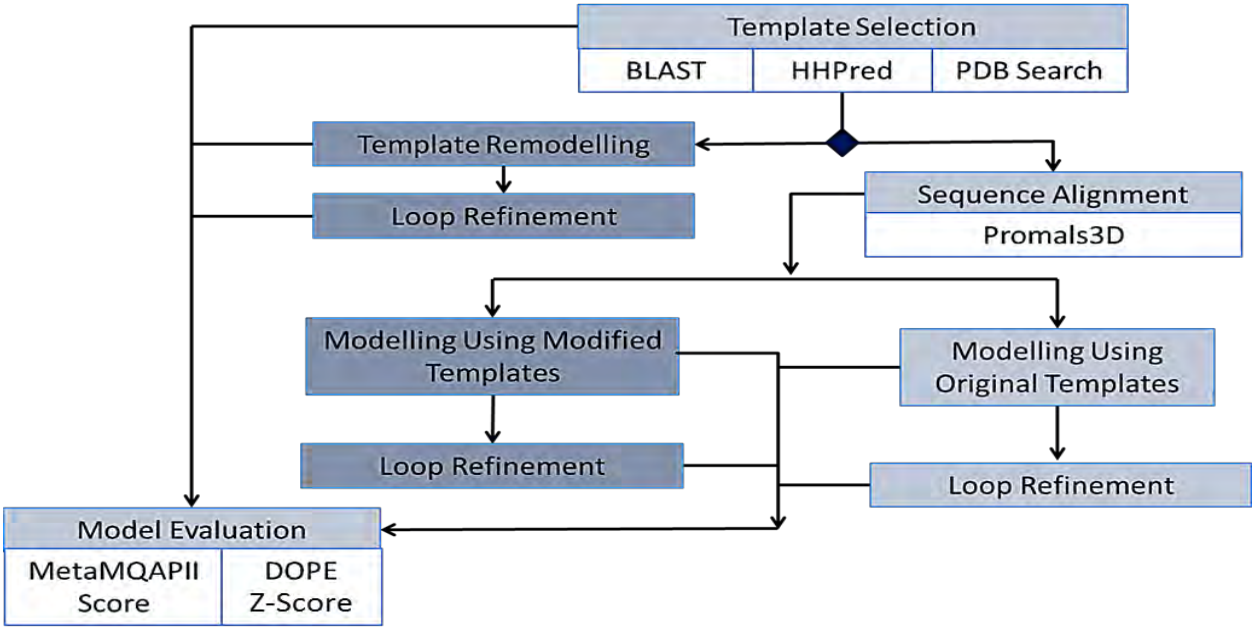

Figure 3.1: Flow diagram describing the homology modelling process. Templates for modelling were identified using BLAST, HHpred and the PDB search tool. Templates were then remodelled and loop-refined. The PfHsp70 protein sequences were aligned with the templates using PROMALS3D and then modelled based on the different templates identified. Modelling was performed using the original templates and the remodelled templates and the results were compared. All models produced were analysed using DOPE Z-score and MetaMQAPII.

3.2.1 TEMPLATE SELECTION AND ANALYSIS

Homologous proteins with known structures were searched for using HHpred (Söding *et al.*, 2005), PSI-BLAST (Altschul *et al.*, 1997) and the PDB. Initially, the HHpred online server was used to search for suitable templates by submitting sequences of PfHsp70-1, PfHsp70-2, PfHsp70-3 and PfHsp70-x separately to the server using the default search parameters provided by the server. Additionally, for each sequence, PSI-BLAST was run with a maximum of 8 iterations and only sequences associated with known structures were considered. Finally, the PDB was searched using the site's sequence search tool in order to find additional sequences. Structures were chosen based on sequence identity, coverage of the query sequence, as well as the quality of the structure itself. Once potential templates were acquired, they were visualised using PyMOL (Schrödinger, LLC) and analysed using MetaMQAPII (Pawlowski *et al.*, 2008). Templates chosen include the PDB entries 1YUW, 2KHO and 3D2F. For more info about these structures, refer to results. In order to make the template structures comparable to the models that

would be produced by MODELLER (Sali & Blundell, 1993; Sali, 2009), each template was altered by running the python script, **model.write.py** (refer to Appendix E). This removed all atoms that were not part of the protein itself from the coordinate file of each template, while the coordinates of each residue atom were unchanged. Template 3D2F coordinate file contained 4 separate chains and therefore had to be further altered in order for it to be analysed by MetaMQAPII. The script, **chain_get.py**, was executed in order to remove all but the first chain, which was used as a template for modelling. Finally, the DOPE Z-score for each template was determined (refer to script **calculate_dope.py**).

3.2.2 REMODELLING OF TEMPLATE STRUCTURES

As a part of validation of the homology modelling approach used, the first step taken was to test the effect MODELLER itself had on the modelling process. More specifically, how accurately would MODELLER model a protein if it had a template with 100% sequence identity? In order to test this, each template structure was remodelled by homology modelling, using itself as a template. This process is summarised in Figure 3.2. For the purposes of comparison, the experimentally solved structures will from here on be referred to as the "original templates", whereas the remodelled forms of these templates will be referred to as "modified templates". The template structures 1YUW, 2KHO and 3D2F were each remodelled using MODELLER. This was first done without refinement and then again separately using slow refinement (refer to **model.py**). For unrefined modelling, 500 models were produced for each template, whereas 100 models were produced when modelling with slow refinement. The models were assessed by calculating DOPE Z-score and the best scoring models from each modelling set were submitted to MetaMQAPII for further model quality assessment. Models from the top modelling set, as determined by this assessment, were submitted for loop refinement. The loop refinement process involved identifying problematic loops in each model by reviewing the MetaMQAPII results of each model using PyMOL. These loops were further refined using MODELLER, performed separately using the loopmodel and dope_loopmodel classes (refer to **loop_ref.py**). For each of these modelling sets, 100 models were produced. Once again, each model produced was assessed using the Dope Z-score and the best scoring models were submitted to MetaMQAPII.

3.2.3 PfHsp70 Modelling

The template remodelling resulted in higher quality structures, as determined by and MetaMQAPII and by calculating the DOPE Z-score for these models. As such the modelling of each PfHsp70 protein was performed using the modified templates in addition to the originals, in order to see which approach produced better quality models. This also served as a further validation step as it gave an indication of whether models produced by MODELLER were of sufficient quality to be used as templates for modelling. The initial rational for using the modified templates was because they were of higher quality than the original models. Unfortunately, remodelling of template 1YUW resulted in a lower quality structure. As a result, the modified template selected for 1YUW had only undergone loop refinement, rather than the full remodelling process. Sequence alignments were done using PROMALS3D; aligning the six PfHsp70 proteins with the 3 templates identified (refer to Figure F.1 and F.2). PROMALS3D was used as it uses structural information to guide the alignment process. As such it should produce alignments that are more suitable to homology modelling than other alignment programs might. The reason all six PfHsp70 proteins were included in the alignment is that the modelling process was started long before it was decided that PfHsp70-y and PfHsp70-z wouldn't be included as Hsp70s in this research. The MSA was used to create PIR files for each template-PfHsp70 combination used for modelling. This was done by running the scripts, **pir_maker.py**, followed by **gap_remove_PIR.py**. The header for each PIR file was constructed manually. The same modelling protocol was followed as that used during template remodelling. Modelling was carried out separately using the original and modified templates, with 100 models being produced with slow refinement. Loop modelling was performed using the dope_loopmodel class, as this proved to yield better models when assessed during template remodelling.

3.2.4 Model Validation

As previously stated, all models produced were assessed by MetaMQAPII and by calculating their DOPE Z-scores. The final set of models was further evaluated by ProSA and Verify3D. ProSA scores were calculated by submitting models to ProSA. Verify3D scores were calculated by submitting models to the Genesilico.

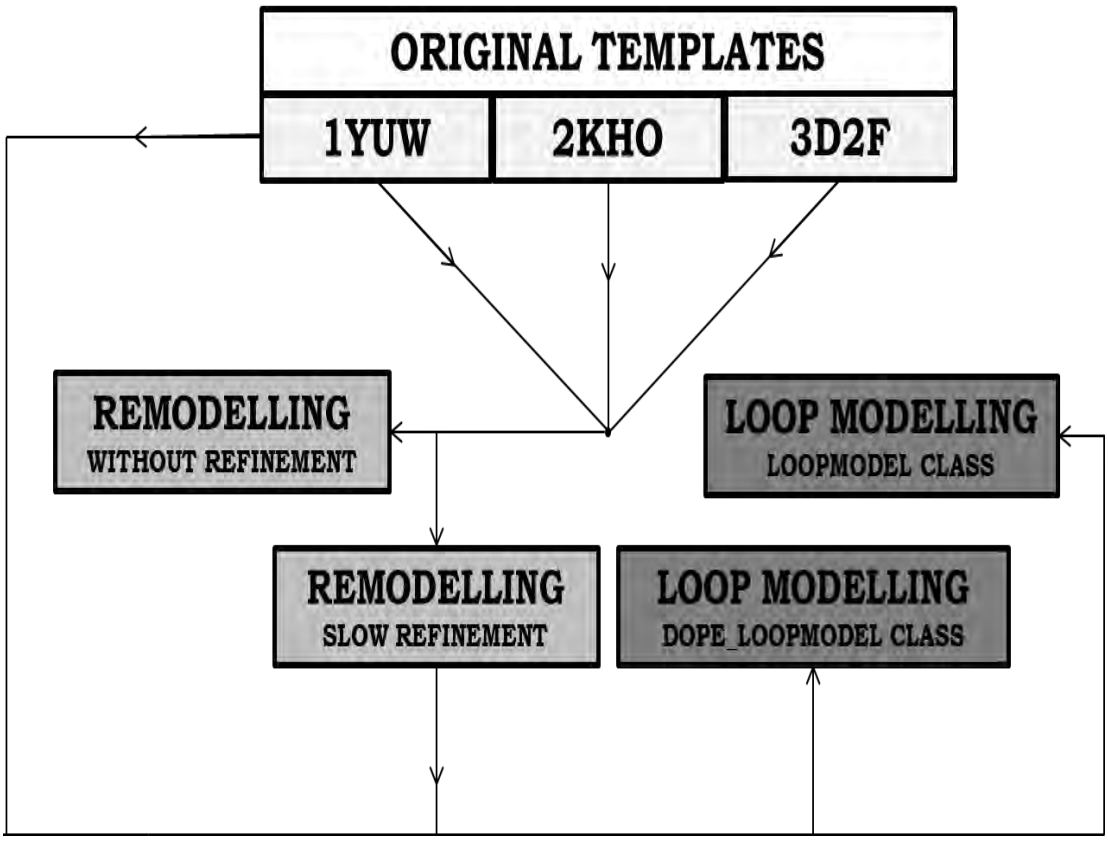

Figure 3.2: Flow diagram describing the process followed for template remodelling. Experimentally solved protein structures 1YUW, 2KHO and 3D2F were each used as templates to model their own 3D structures, using MODELLER. This was done both with and without refinement. The refined structures were then submitted for loop refinement, using two different loop modelling methods available to MODELLER. The original template structure 1YUW was also submitted for these two forms of loop refinement.

3.2.5 Analysis of Sequence Features

The top model of each PfHsp70 protein in each conformation was selected based on the results of the model validation process mentioned above. In order to ensure that residues could be accurately mapped to the models produced, the PDB file of each model was renumbered using the script **renum.py**. This ensured that the residues in the PDB file were numbered according to their positions in the protein. The position of each sequence feature mentioned in Chapter 2 was recorded. BioEdit was used to determine the position of each residue. Sequence features were then mapped to each PfHsp70 in each conformation using PyMOL. Residues were analysed by inspection

3.3 RESULTS

3.3.1 TEMPLATE SELECTION

The sequences of the PfHsp70-1, PfHsp70-2, PfHsp70-3 and PfHsp70-x were each submitted to PSI-BLAST and HHpred in order to identify potential templates. Additionally, the PDB was searched for potential templates that may have been missed by other methods. The search by HHpred returned the same three high-scoring templates for each of the proteins. Across all four PfHsp70 sequences, these three templates were always within the top 4 search results. All three of these were considered, as each was in a different conformation, therefore potentially representing the protein at a different stage of the Hsp70 ATPase cycle. Results returned by PSI-BLAST and the manual search of the PDB failed to identify more suitable templates. Details about each template are shown in Table 3.1.

Table 3.1: Information about templates selected for modelling. This includes the PDB ID of each template, the protein the template represents, the method used to solve its structure and information relating to the quality of the template. The GDT_TS score is reported as predicted by MetaMQAPII.

	PDB ID	Chain	Protein	Organism	Method used to solve structure	Resolution (Å)	Predicted GDT_TS
Template 1	1YUW	**A**	Hsc70	*Bos taurus*	X-Ray Diffraction	2.6	78.07
Template 2	2KHO	**A**	DnaK	*Escherichia coli*	Solution NMR	N/A	69.58
Template 3	3D2F	**A**	Sse1	*Saccharomyces cerevisiae*	X-Ray Diffraction	2.3	84.54

3.3.2 TEMPLATE REMODELLING

3.3.2.1 ANALYSIS OF THE MODELLING PROCESS

Each template 1YUW, 2KHO and 3D2F was remodelled, using its own structure as a template (refer to Figure 3.2). Quality assessment scores of the top models produced during each step of the process are given in Table 3.2. The remodelling of template 1YUW produced models with both an increase in the DOPE Z-score and a decrease in the predicted GDT_TS score. Remodelling of 2KHO and 3D2F improved the quality of the templates with respect to both

these scores. Also seen in Table 3.2, was that using refinement protocol when modelling lead to the production of models with a higher DOPE Z-score when compared to modelling without refinement. Loop refinement protocol used resulted in an increase in the DOPE Z-score score, with a small increase in the predicted GDT_TS score. When comparing loop refinement using the loopmodel class with that using the dope_loopmodel class, the dope_loopmodel class produced models with an equal or higher DOPE Z-score score.

Table 3.2: Summary of the top models produced during template remodelling. Each of three templates (PDB IDs: 1YUW, 2KHO and 3D2F) were remodelled using different options provided by MODELLER. Modelling was carried out either without refinement (w/r) or using slow refinement (s/r). Loop modelling was performed to improve the loops within each top-scoring remodelled template. This was done using either the **loopmodel** (l) or **dope_loopmodel** (d/l) option. Applicable only to the 1YUW models, loop modelling was performed on both the original template (O) and the remodelled template (M). The DOPE Z-score of each model, as well as the GDT_TS score as predicted by MetaMQAPII is given.

Model	DOPE Z-score	Predicted GDT_TS
Template 1YUW		
Original Template	-0.94	78.07
Modelling (w/r)	-0.84	76.26
Modelling (s/r)	-0.88	75.99
Loop Modelling (l) (M)	-0.89	78.20
Loop Modelling (l) (O)	-0.96	78.88
Loop Modelling (d/l) (M)	-0.90	76.08
Loop Modelling (d/l) (O)	-0.96	78.79
Template 2KHO		
Original Template	-0.77	69.58
Modelling (w/r)	-0.77	69.38
Modelling (s/r)	-0.85	69.92
Loop Modelling (l)	-0.89	70.67
Loop Modelling (d/l)	-0.93	70.96
Template 3D2F		
Original Template	-1.20	84.54
Modelling (w/r)	-1.23	85.14
Modelling (s/r)	-1.28	85.85
Loop Modelling (l)	-1.27	85.61
Loop Modelling (d/l)	-1.32	86.05

Table 3.3: Top PfHsp70 protein models. Models were evaluated by their DOPE Z-score, predicted GDT_TS scores, Global RMSDs, ProSA Z-scores and Verify3D scores. Rows highlighted in grey represent models that were built based on modified templates. Rows that are not highlighted represent models built using the original templates.

Model	DOPE Z-score	GDT_TS	RMSD	ProSA Z-score	Verify3D
PfHsp70-1_1YUW.pdb	-0.90	79.64	2.01	-10.15	236.31
PfHsp70-1_2KHO.pdb	-0.68	64.06	2.43	-11.04	240.08
PfHsp70-1_3D2F.pdb	-1.06	76.68	1.77	-11.59	255.80
PfHsp70-2_1YUW.pdb	-1.00	74.73	1.88	-10.30	249.78
PfHsp70-2_2KHO.pdb	-0.79	69.52	2.11	-11.19	265.75
PfHsp70-2_3D2F.pdb	-1.01	75.98	1.89	-11.13	261.88
PfHsp70-3_1YUW.pdb	-0.78	76.37	1.88	-9.38	242.92
PfHsp70-3_2KHO.pdb	-0.82	71.89	1.90	-10.55	254.26
PfHsp70-3_3D2F.pdb	-0.96	73.72	1.90	-10.71	261.44
PfHsp70-x_1YUW.pdb	-0.91	78.83	1.89	-10.41	257.58
PfHsp70-x_2KHO.pdb	-0.77	67.64	2.14	-11.59	260.91
PfHsp70-x_3D2F.pdb	-1.09	76.66	1.66	-11.94	264.55

3.3.3 PfHsp70 Protein Modelling

Using the information gathered from the template remodelling, a set protocol was developed for modelling using MODELLER. Modelling of each PfHsp70 protein was performed using both the modified templates and the original templates. Each protein was modelled using each of the 3 different templates. Of the 4800 models that were produced, 144 were evaluated (The top 3 from each modelling set) using MetaMQAPII, including comparisons of local quality by visualising the MetaMQAPII results in PyMOL. By combining the MetaMQAPII and DOPE Z-scores calculations of each of these models, a top model was selected for each protein, as modelled using either an original or modified template. Evaluation results of the top models are shown in Table 3.3. As indicated, six of these were models produced using modified templates and the other six with the originals. All proteins modelled using template 2KHO were modelled better using the modified templates, rather than the originals (refer to Appendix G for comparisons). In addition to the scores given in Table 3.3, MetaMQAPII also provides a per-residue score, which reflects the predicted quality of the residue. This is score can be used to assess the local quality of a model and determine problem areas. Below in Figures 3.3 - 3.6 are the top models of each

PfHsp70, along with the template used to model it. These models are coloured according to the per-residue scores assigned by MetaMQAPII. The spectrum for this includes colours from blue to red, with dark blue indicating good quality and red indicating problematic residues. As seen in Figures 3.3 and 3.4, the structures modelled using template 1YUW are all good quality models with few problematic areas. The models of proteins modelled using template 2KHO (Figure 3.5) contain more regions of poor local quality when compared to the original template. The models shown in Figure 3.6 appear as good quality structures.

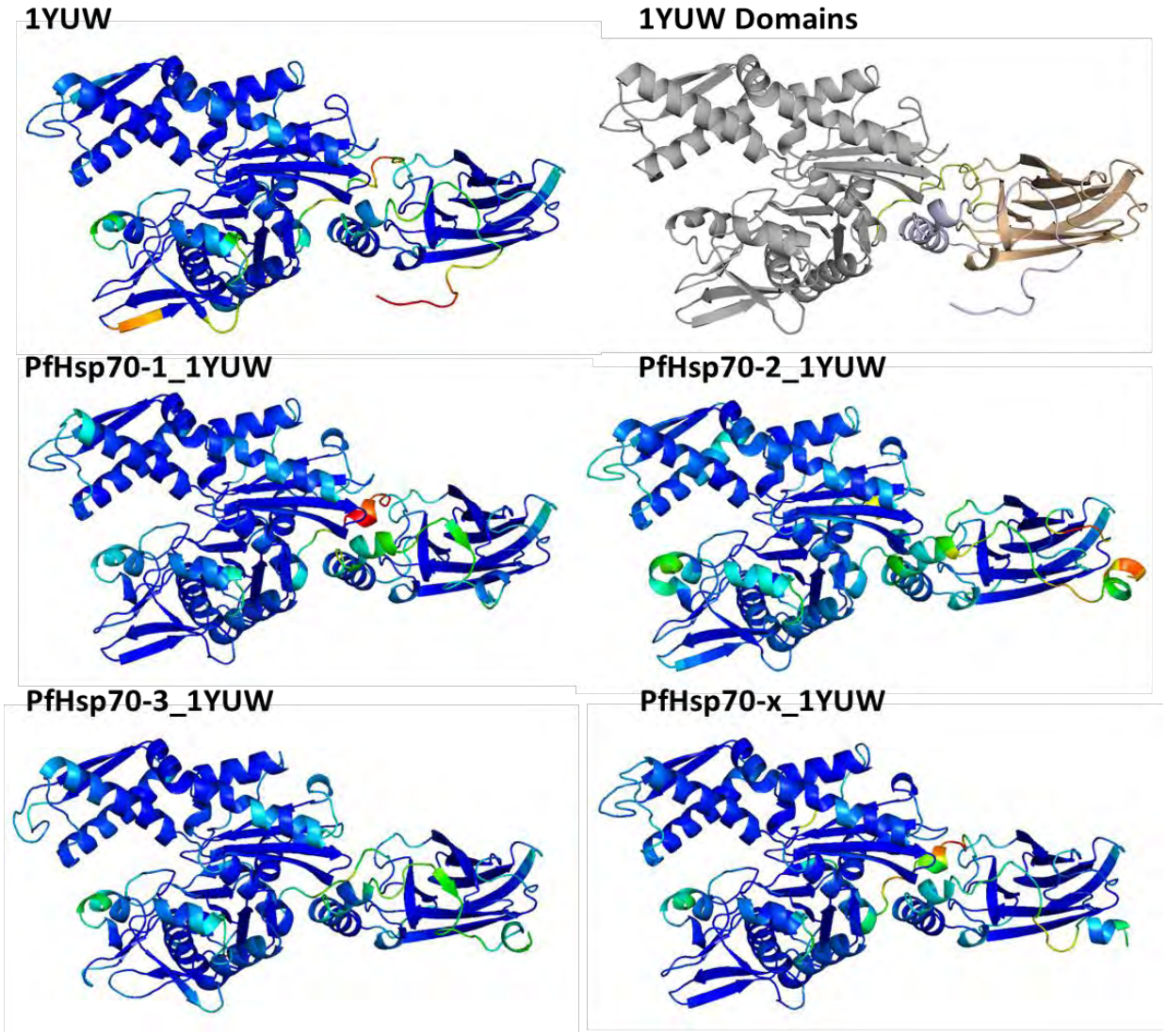

Figure 3.3: PfHsp70 proteins, modelled using template 1YUW emphasising the quality of the ATPase domain. Structures were visualized using PyMOL and are coloured according to their MetaMQAPII local quality scores. The 'Domains' model is a reference model where the domains are coloured as follows – Grey: ATPase Domain; Lemon: Linker Region; Wheat: SBD; Pale Blue: Alpha-Helical Lid.

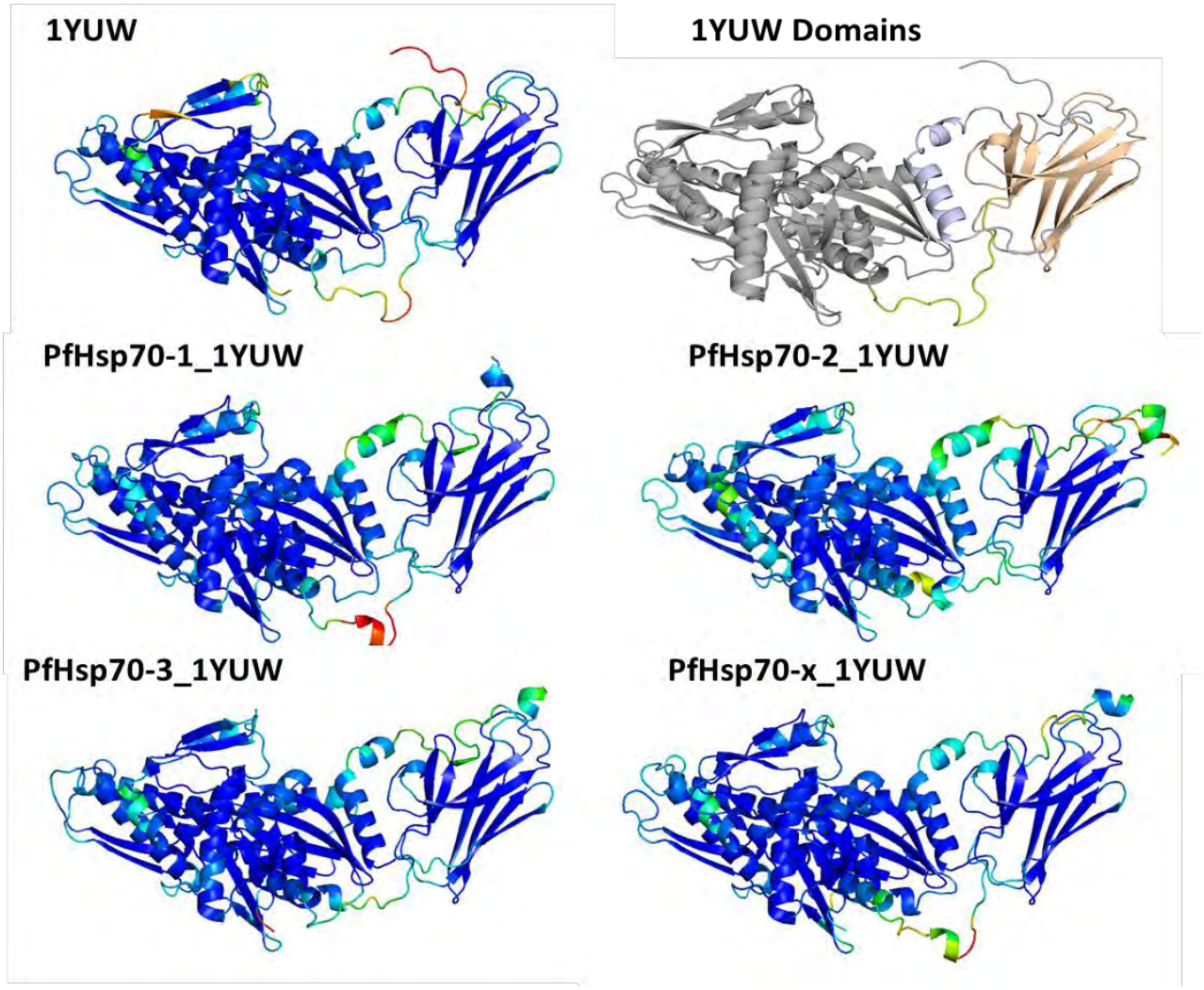

Figure 3.4: PfHsp70 proteins, modelled using template 1YUW, emphasise the quality of the SBD. Structures were visualized and coloured as in Figure 3.3.

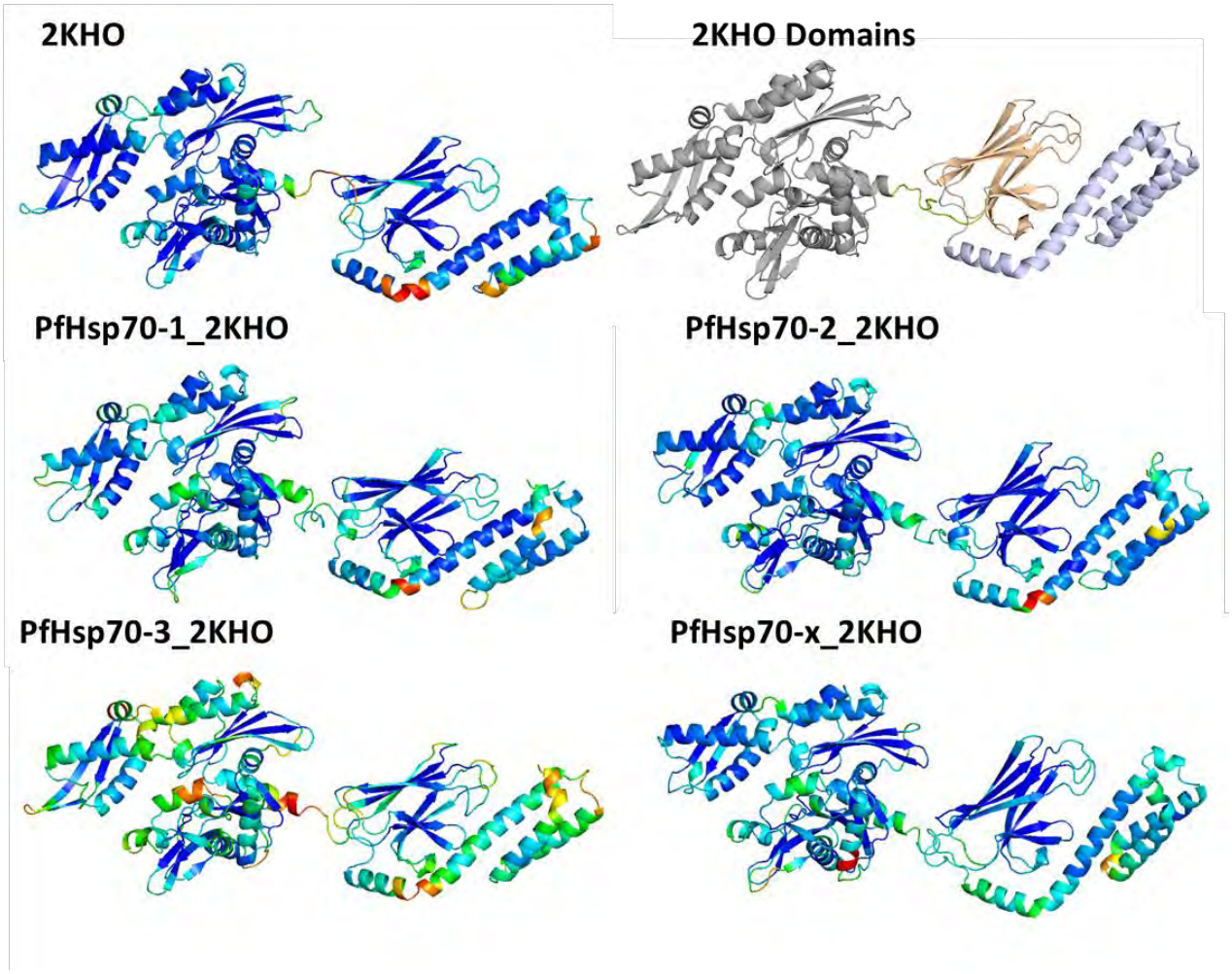

Figure 3.5: PfHsp70 proteins, modelled using template 2KHO. Structures were visualized and coloured as in Figure 3.3

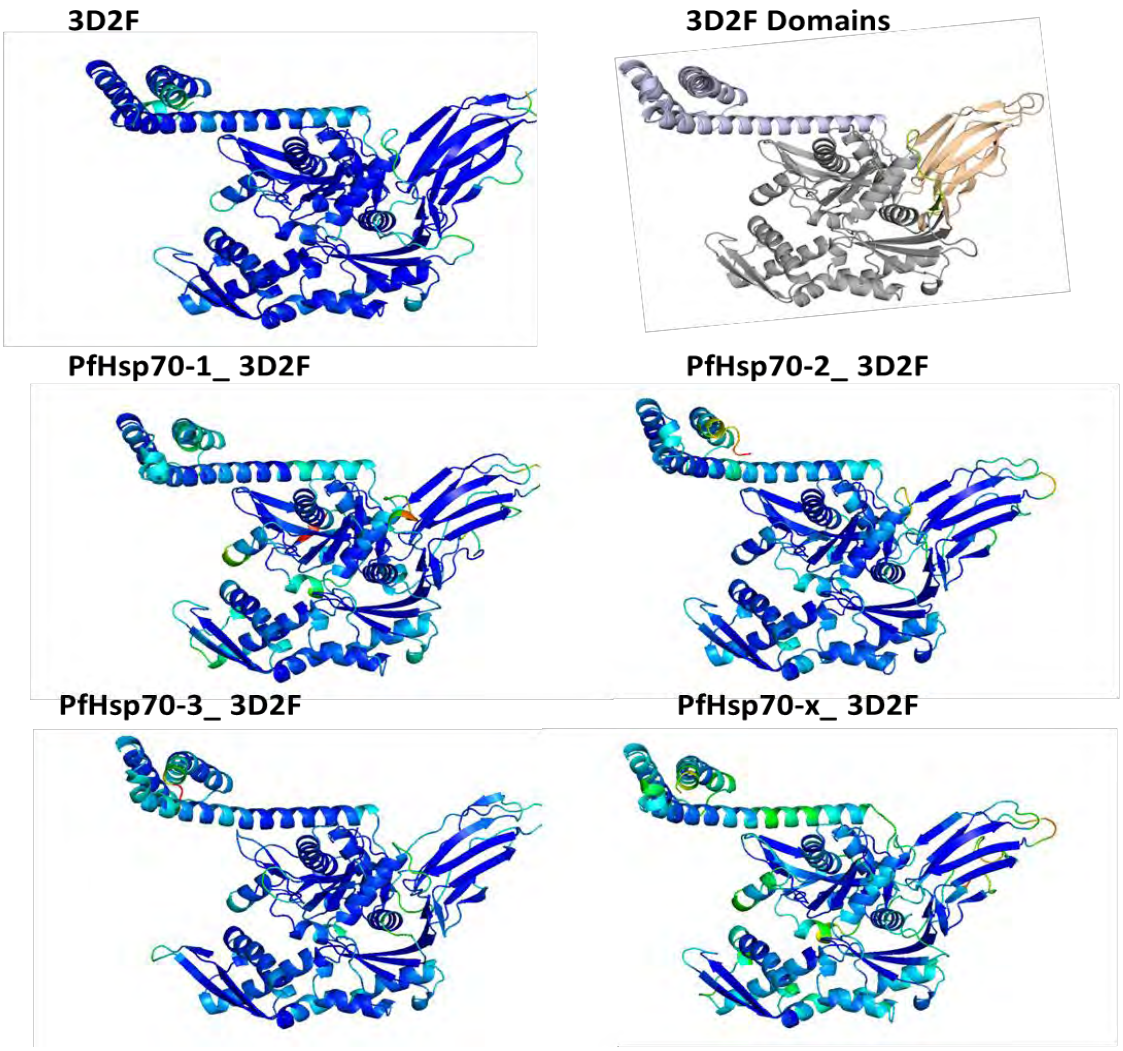

Figure 3.6: PfHsp70 proteins, modelled using template 3D2F. Structures were visualized and coloured as in Figure 3.3

3.3.4 Analysis of Sequence Features

Sequence features that were determined in Chapter 2 were mapped to the PfHsp70 models. Two sets of models were produced, based on the two types of sequence analysis performed. The first was an interaction set, which focused on residues which were predicted to interact with other proteins, including peptide substrates. The second set was based on sequence features specific to the different types of Hsp70s analysed. Where models are presented, only one PfHsp70 protein is given in each figure. This is mainly due to space considerations, but also because there is very difference between most sequence features as they appear across the different PfHsp70s considered. These sequence features are, however, marked on the models in Appendixes H and I.

3.3.4.1 J-Domain Contact Residues

Residues believed to bind to the J-domain of Hsp40-like proteins were mapped to all models produced. Shown in Figures 3.7-3.9 below, are these residues mapped to PfHsp70-2. These represent the protein in different conformations, depending on the template used during modelling. Figure 3.7 represents PfHsp70-2 as modelled based on template 1YUW. Here the ATPase domain, SBD and alpha-helical lid are in close proximity to one another and as a result, the J-domain biding residues are partially closed off. This is most notable with Val 238 of PfHsp70-2, which is completely covered by the alpha-helical lid. A similar effect is seen with the other PfHsp70s in this conformation (refer to Appendix H). In Figure 3.8, PfHsp70-2 is modelled based on template 2KHO, which represents the protein when the SBD and alpha-helical lid are more distanced from the ATPase domain. In this conformation, the J-domain contact residues are more accessible. In conformation of 3D2F (Figure 3.9), The SBA and ATPase domain are in close proximity to each other, with the alpha helical lid facing away from the SBD. Here the binding region is again more open than in Figure 3.7; however, binding to this region may be partially inhibited by the SBD which may close partially over Asp198 of PfHsp70-2.

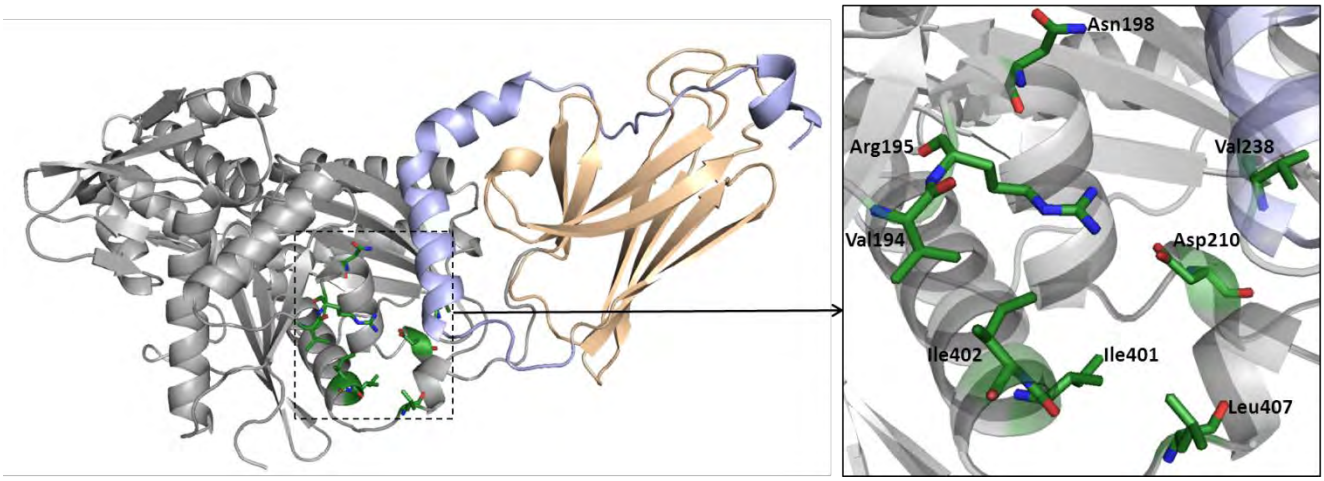

Figure 3.7: J-domain contact residues of PfHsp70-2 in the conformation of template 1YUW. The protein PfHsp70-2, as modelled based on template 1YUW is shown on the left with putative J-domain binding region indicated by a dashed box. A close-up view of this region

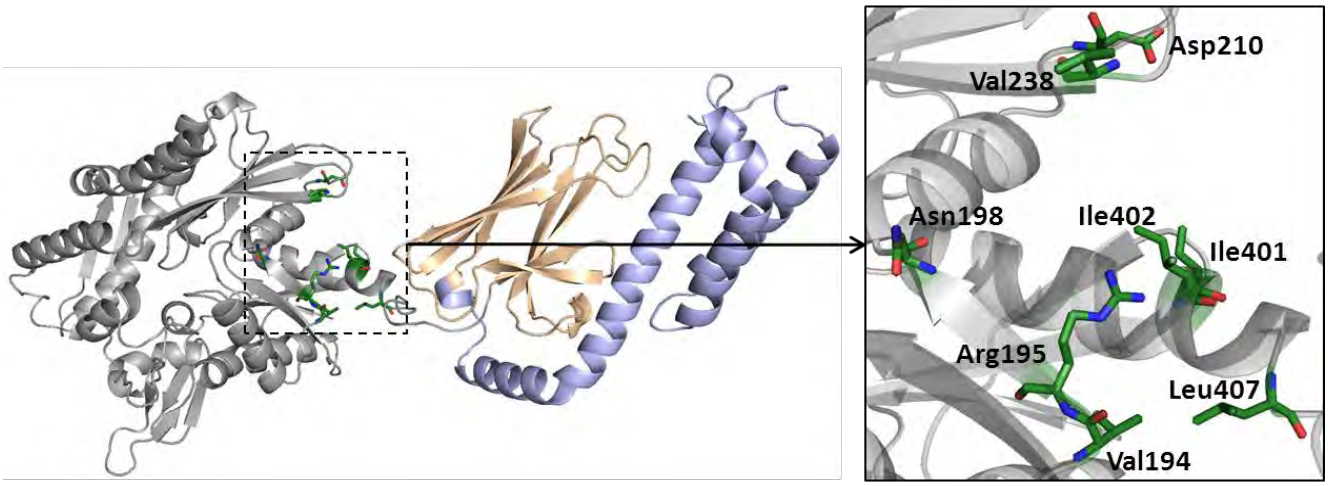

Figure 3.8: J-domain contact residues of PfHsp70-2 in the conformation of template 2KHO.
This figure is labelled and coloured as per Figure 3.7.

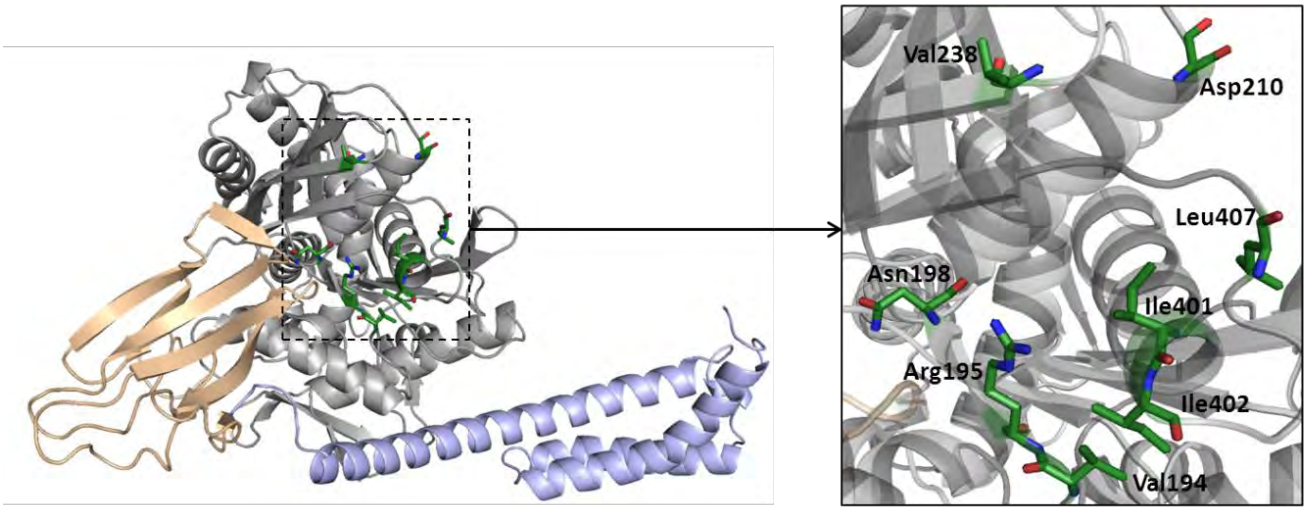

Figure 3.9: J-domain contact residues of PfHsp70-2 in the conformation of template 3D2F.
This figure is labelled and coloured as per Figure 3.7.

3.3.4.2 SBD Contact Residues

Three different types of residues, believed to be important to peptide binding were mapped to the SBD of each model. Figures 3.10 – 3.12 represent PfHsp70-x, as modelled based on templates 1YUW, 2KHO and 3D2F, respectively. As can be seen in Figure 3.10, in the conformation of template 1YUW, residues of the alpha-helical are bound by the SBD. Here the highlighted residues are all positioned around the helical lid, close enough to potentially interact with its residues (data not shown). In Figure 3.11, the alpha helical lid is folded over the SBD and the highlighted residues adopt a similar conformation to those shown in Figure 3.10. When the lid extends away from the SBD, as seen in Figure 3.12, the highlighted residues close in on each other and close the substrate binding pocket.

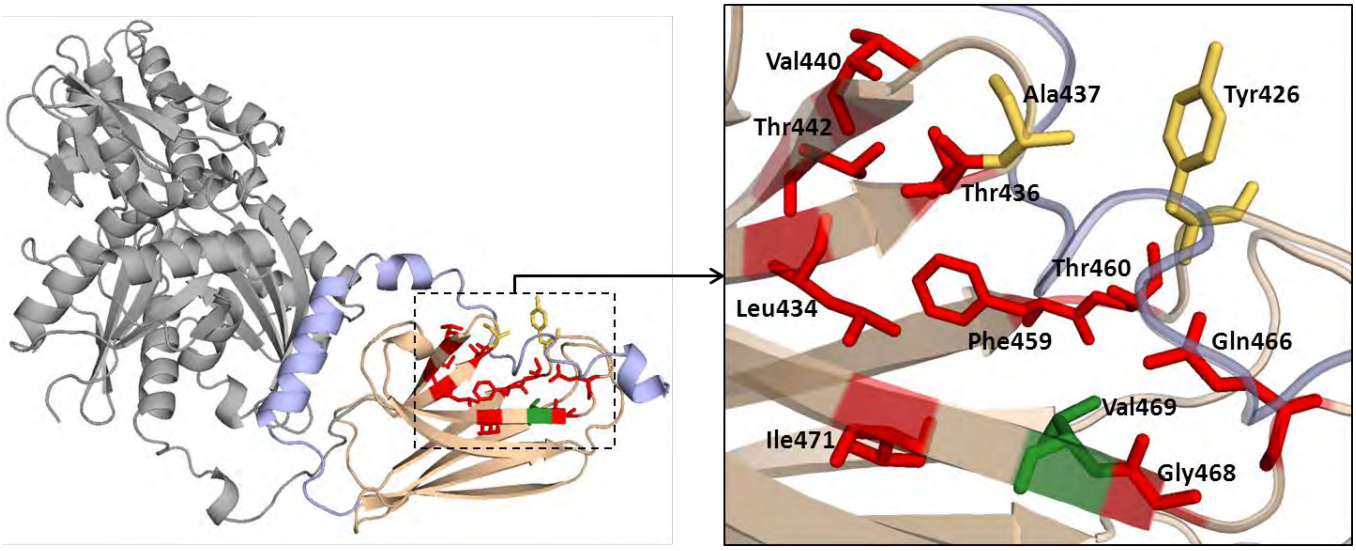

Figure 3.10: SBD contact residues of PfHsp70-x in the conformation of template 1YUW. The protein PfHsp70-x, as modelled based on template 1YUW is shown on the left. Peptide binding occurs within the area indicated by a dashed box. A close-up view of this region is shown on the right, with residues believed to be important in peptide binding shown in stick view. Residues are coloured as follows: Red – residues which make contact with the substrate during peptide binding. Yellow – Hydrophobic arch residues. Green – Residue which forms the hydrophobic pocket. The rest of the protein is coloured as follows: Grey - ATPase domain; Wheat – SBD; Pale Blue – Alpha helical Lid.

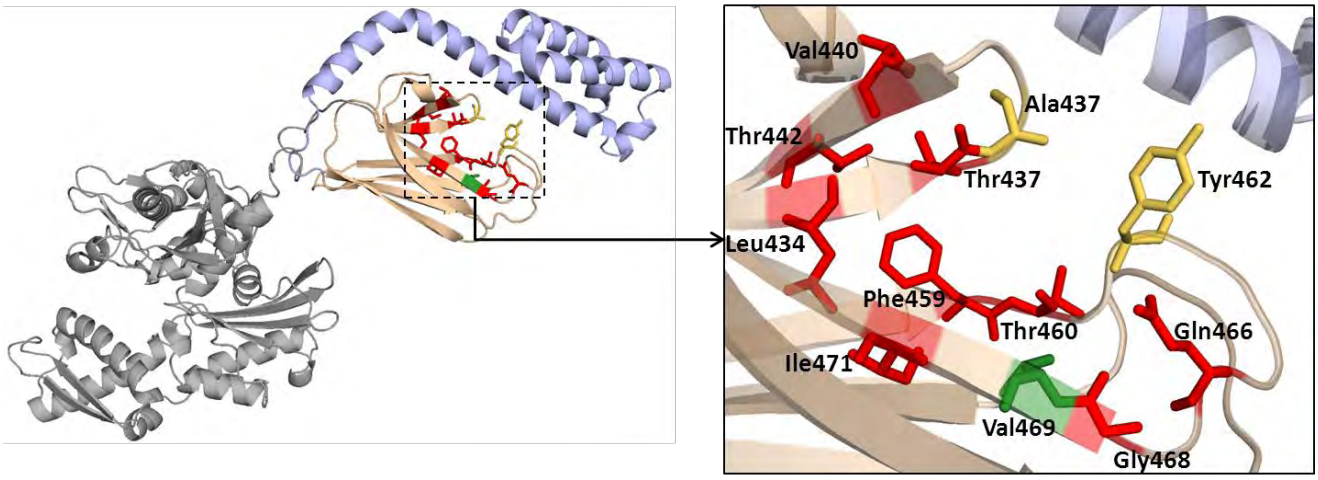

Figure 3.11: SBD contact residues of PfHsp70-x in the conformation of template 2KHO. This figure is labelled and coloured as per Figure 3.10

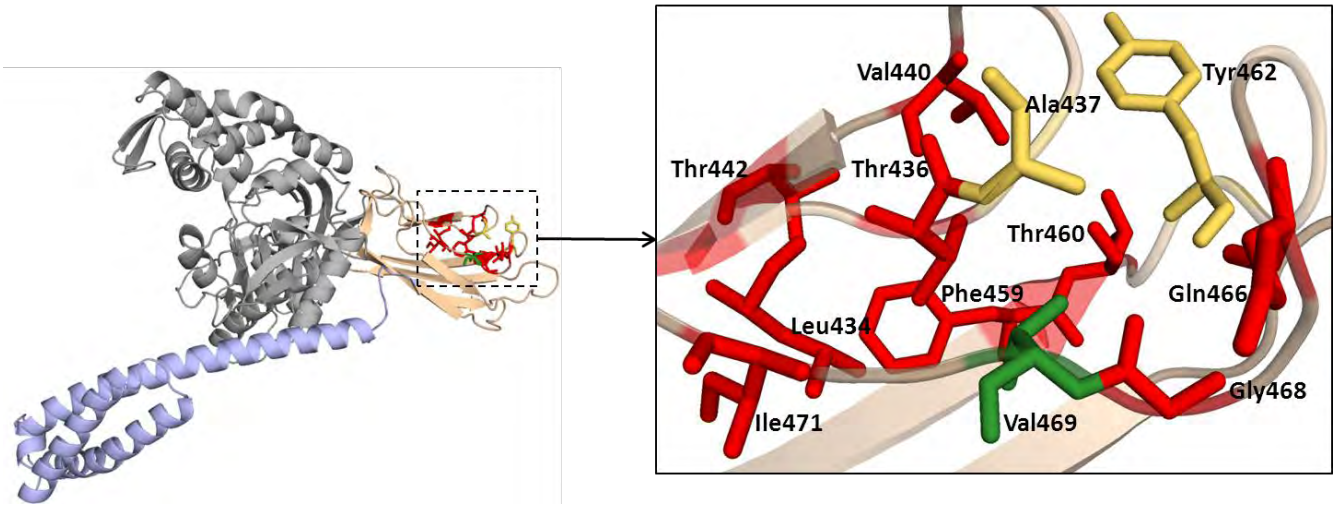

Figure 3.12: SBD contact residues of PfHsp70-x in the conformation of template 3D2F. This figure is labelled and coloured as per Figure 3.10

3.3.4.3 Bag Domain and Mge1p Contact Residues

Figures 3.13 – 3.15, below indicate the predicted Bag domain binding residues of PfHsp70-1, as seen in three different conformations on the protein. In all Figures it can be seen that the conformation of the ATPase domain changes very little as it interacts differently with the other domains. As a result, the Bag domain binding residues are always open for interaction with the potential NEF. A possible exception to this is represented by the conformation of 3D2F (Figure 3.15), where the alpha-helical lid may prevent interaction with Asp57 of PfHsp70-1. This effect is also seen with the Mge1p binding residues of PfHsp70-3 (refer to Appendix H).

3.3.4.4 Hsp110 Contact Residues

As with Bag domain and Mge1p binding residues, the residues believed to interact with Hsp110 during nucleotide exchange were mapped to each PfHsp70 protein, with the exception of PfHsp70-3 (refer to Appendix H). These were all found to be surface residues and exposed to interaction, regardless of conformation.

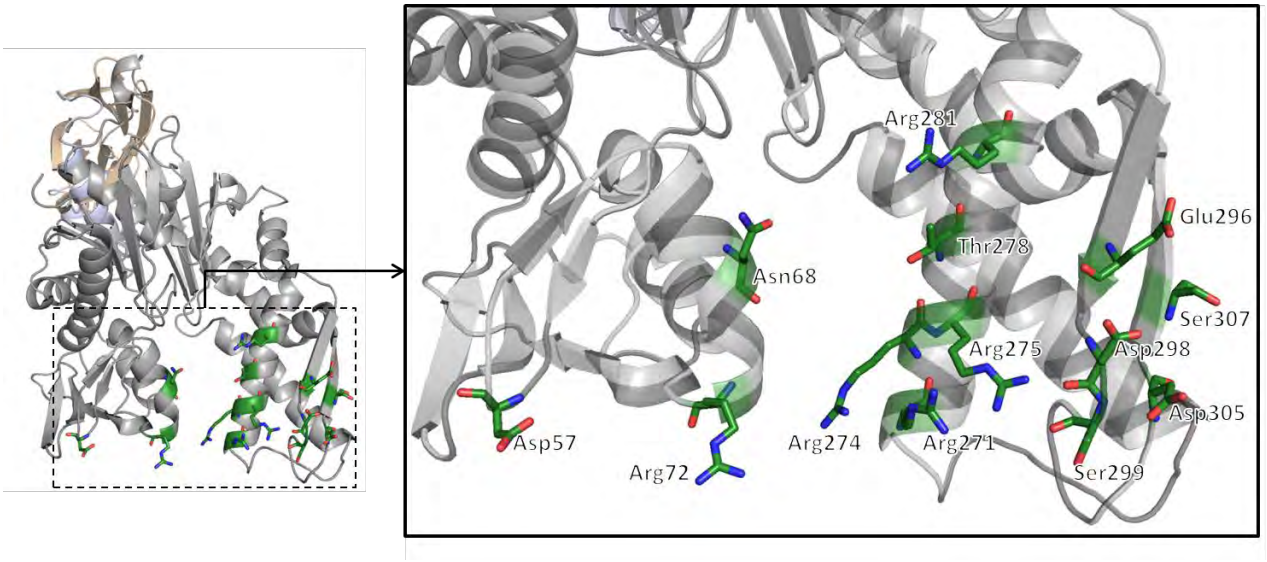

Figure 3.13: Bag domain contact residues of PfHsp70-1 in the conformation of template 1YUW. On the left, PfHsp70-1, as modelled based on template 1YUW with the region of the protein believed to interact with Bag domain proteins indicated by a dashed box. A close-up view of this region is shown on the right; with residues believed bind the bag domain shown in stick view coloured in green, with oxygen and nitrogen atoms coloured red and blue, respectively. The rest of the protein is coloured to indicate the different domains, as follows: Grey: ATPase domain; Wheat: SBD; Pale Blue: Alpha helical Lid.

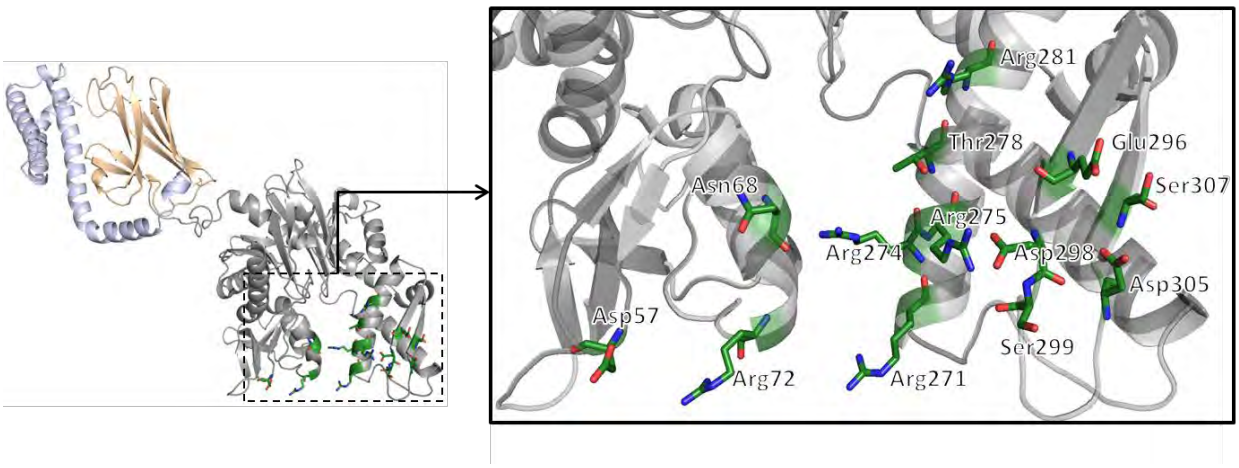

Figure 3.14: Bag domain contact residues of PfHsp70-1 in the conformation of template 2KHO. This figure is labelled and coloured as per Figure 3.13, above.

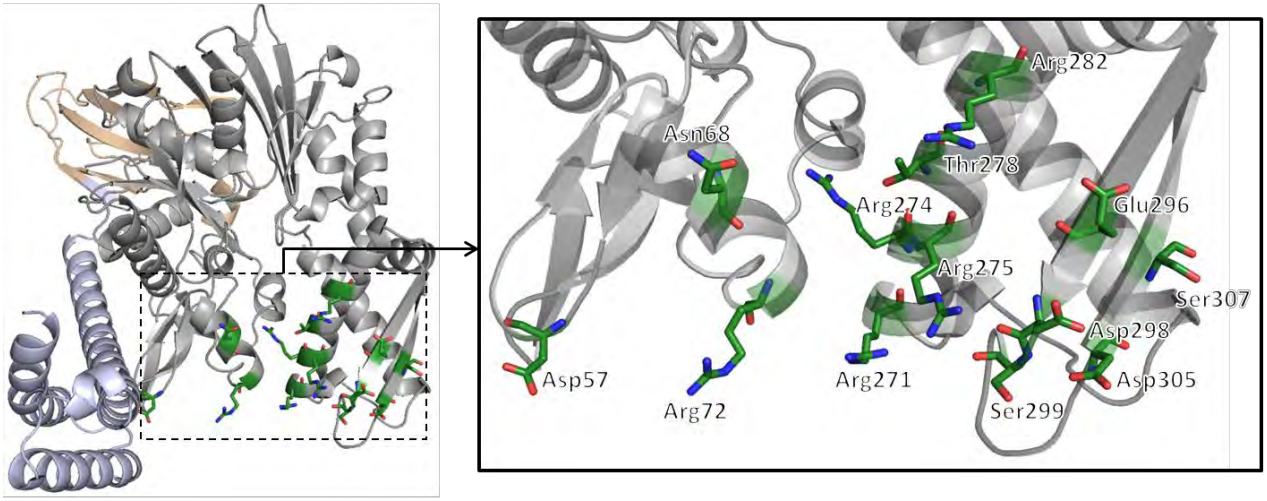

Figure 3.15: Bag domain contact residues of PfHsp70-1 in the conformation of template 3D2F. This figure is labelled and coloured as per Figure 3.13, above.

3.4 DISCUSSION

3.4.1 HOMOLOGY MODELLING PROCESS

3.4.1.1 TEMPLATE SELECTION AND ALIGNMENTS

In terms of template selection, HHpred displayed the most promising results when identifying suitable templates for homology modelling. Although the other two approaches considered were able to identify the same structures, these results were buried among other structures, which were identified as more suitable. Often this was a result of a sequence identity bias, as some of the top structures returned each presented a measured query coverage of approximately 55%, as opposed to the minimum of 80% query coverage of the templates used. The templates chosen included bovine Hsc70, *E. coli* DnaK and *S. cerevisiae* Sse1p. These structures were each solved in a different conformation. Template 1YUW (Jiang *et al.*, 2005), was crystalized with the ATPase and SBD in contact with each other. The structure of template 2KHO (Bertelsen *et al.*, 2009), was solved in an ADP- and substrate bound state, with the ATPase and SBD completely separated from one another. Finally, Sse1p (Template 3D2F, Polier *et al.*, 2008) represents a yeast Hsp110, related to Hsp70s, in an ATP-bound state. The structure of this protein is also fairly more compacted, with the alpha-helical lid closing over the ATPase domain. The second crucial step in homology modelling is to ensure that the template sequence is adequately aligned with the target sequence (Krieger *et al.*, 2003; di Luccio & Koehl, 2011). The PROMALS3D webserver (Pei *et al.*, 2008) was used for sequence alignment because it was developed specifically for protein structure modelling. It incorporates secondary structure prediction methods into its HMMs, in addition to using 3D structural information of the templates used to improve the quality of the alignment. Additionally, it was shown in the previous chapter (Refer to Appendix B) that the PROMALS3D and MAFFT align Hsp70s in almost identical ways across the sequence, even though these two alignment programs use different algorithms and parameters to align sequences. Exceptions are seen in the N- and C-terminal domains, although these weren't covered by the template structures in any case.

3.4.1.2 TEMPLATE REMODELLING

As a validation step, each of the three templates was remodelled using MODELLER. Evaluation of remodelled templates 2KHO and 3D2F revealed that these models scored higher than the original template structures, when evaluated by DOPE Z-score and MetaMQAPII GDT_TS score (Table 3.2). This process also allowed the comparison of different modelling options made available by MODELLER. In all instances of remodelling, the slow refinement method produced models with lower DOPE scores than their unrefined counterparts. The models were further refined by adjusting the loops to bring them to a lower energy state. This was done using either the loop refinement method, or the DOPE-based loop refinement method. The difference between the methods is the energy scores MODELLER uses to adjust the loops (Sali, 2009). With all three templates, the DOPE-based loop refinement resulted in models with equal or lower DOPE scores (Table 3.2). This was expected, however, since the energy function used by this method is the same function used to calculate DOPE score. It is interesting to note that no particular method with regard to both modelling and loop refinement had a comparably favourable effect on the GDT_TS score. Although in most instances this score was improved, it could not be used to separate any of the given modelling strategies. As such, the strategy chosen was one which produced a more favourable DOPE score the most, whilst not lowering the predicted GDT_TS score. Although these scores were used to distinguish between different methods, there was very little difference seen in any given comparison. This chapter also lacks the necessary statistics to determine whether or not these methods are significantly different. This is because these methods are only assessed, based on the top models they produce. In order to generate enough top models for statistical purposes, hundreds of modelling runs would have to be performed, producing tens of thousands of models. It can be said, though, that there is no indication that modelling using MODELLER lowers the quality of models representing a protein structure.

3.4.1.3 USE OF MODIFIED TEMPLATES

Another aspect of modelling addressed in this chapter was the use of the modified templates for homology modelling, as compared to using the original template structures. Modelling of the PfHsp70 proteins was done using both the modified structures as well as the original templates.

Of the 12 top protein models chosen, six of were modelled using the modified templates. Nine of these top models were scored better than their counterparts by most, if not all MQAPs used (Refer to Appendix G). However, three models presented problems in determining which were best. For models PfHsp70-2_2KHO and PfHsp70-x_3D2F, the local scores contradicted the global score quite substantially. In these instances, the global score was considered to be more accurate, because a variety of global scores were considered. Even though MetaMQAPII uses many different quality assessment programs to assign scores to models, it is still just a single program. The other uncertainty rose when comparing the PfHsp70-2_1YUW, modelled using the modified template as compared to its original template counterpart. In this instance, there was no conclusive evidence given by the global scores that one of the models was best and as a result the local score was assessed, by viewing the MetaMQAPII scores in PyMOL, in order to choose a top model. Although the approach of modelling using remodelled templates does show promise, it also needs to be validated, by statistical analysis, using a wider range of templates of varying quality. This is discussed in greater detail in Chapter 4. It is possible that the approach is suited only to lower quality template structures. This is based on the results of models produced using template 2KHO, wherein the modified templates consistently outperformed the originals. These results do at least validate the modelling process of MODELLER. The remodelled templates, which were derived using an *in silico* technique were of high enough quality to produce models comparable to models produced using the original templates.

3.4.2 Analysis of Sequence Features

In the previous chapter, various sets of important residues were determined by studying sequence features identified in other Hsp70s. These residues were then highlighted in the set of proteins modelled. The primary sequence features highlighted were those important to the ATPase cycle of Hsp70s. Since these residues interact with other proteins, it is important that they are exposed on the surface of the protein, so they are available for interaction. Another interesting aspect studied was whether the conformation of the protein affected the arrangement of these residues and their accessibility to other proteins. The mapping of residues to the models of each PfHsp70 considered is presented in Figures 3.7 – 3.15. A representative PfHsp70 protein is shown per set of interacting residues. These residues were, however, mapped to all relevant PfHsp70 proteins and can be viewed in Appendix H.

3.4.2.1 J-DOMAIN CONTACT RESIDUES

All three types of Hsp70s highlighted in this research are known to interact with some form of J-domain protein. The residues involved in this interaction were found to be located on the ATPase domain of all four canonical PfHsp70s (Figures 3.7 - 3.9; Appendix H). Highlighted in Figures 3.7-3.9 was PfHsp70-2. This protein was shown because, like PfHsp70-3, one of the known putative J-domain contact residues was missing from this protein. Study of this deletion at a structural level revealed that the position of this residue on the structure was replaced by next immediate J-domain contact residue. In other words, at a structural level, the Ser398 of PfHsp70-1 is replaced by a leucine in both PfHsp70-2 and PfHsp70-3. It is possible that the final J-domain contact residue of PfHsp70-1 (Val401), is either not necessary for interaction with the J-domain or that this contact point is present as one of the valine or leucine residue further down the linker region in PfHsp70-2 and PfHsp70-3. It was surprising to find that region of the ATPase involved in J-domain binding was most exposed in the conformation adopted by template 2KHO. This conformation was captured when the protein was already bound to substrate and ATP had already been hydrolysed to ADP, and therefore contact with a J-domain protein shouldn't be necessary. However, this binding region was largely exposed in all conformations studied. To gain an adequate understanding of this binding process, these proteins would probably need to be modelled with their specific J-domain proteins.

3.4.2.2 SBD CONTACT RESIDUES

Although the conformations of the proteins modelled by templates 1YUW and 2KHO (Figures 3.10 and 3.11, respectively) were very different, analysis of the substrate binding residues indicated that in the both these conformations the protein was binding a peptide substrate. This can be compared to the proteins modelled with template 3D2F (Figure 3.12), wherein the hydrophobic pocket does not form when the α-helical lid is orientated away from the SBD. There has been much work done regarding the mechanism of substrate binding in Hsp70/DnaK and the α-helical lid has been found to be important in stabilising substrate binding (Moro *et al.,* 2004). It has been shown by Pellecchia *et al.* (2000) that the removal of the α-helical lid from DnaK does not abolish substrate binding or ATP-induced release of substrate. However, research by Mayer *et al.* (2000) revealed that without the α-helical lid, DnaK is unable to refold denatured protein

substrates. It has additionally been shown that the lid of DnaK will open to varying degrees to accommodate different types of substrate and will often not enclose the peptide (Schlecht *et al.*, 2011). This indicates that the lid doesn't stabilise substrate binding by closing over the peptide substrate. The work presented in this project suggests that the residues of the SBD form a substrate binding pocket, based on the position of the alpha-helical lid, relative to the SBD. This would indicate that the role of the α-helical lid in substrate binding is an interaction with the β subdomain of the SBD, rather than the substrate itself.

3.4.2.3 NEF CONTACT RESIDUES

Residues involved in the interaction with NEFs were open and available for interaction regardless of the conformation of the protein (Figures 3.13-3.15, Appendix H). This was most noticeable with residues involved in Bag domain binding and Mge1p binding. The residues involved in Hsp110 binding covered most of an entire side of the protein. Although the conformation of these residues on the protein changes with conformation, analysis of these residues in NEF binding would require the NEF, itself, in order to determine which conformation is most viable for binding. It does appear, though, that the availability of these residues for binding is not the only factor which influences whether or not an NEF will bind to Hsp70.

3.4.3 CONSERVED RESIDUES OF PFHSP70 PROTEINS

The residues identified in the previous chapter as being conserved in eukaryotes, Apicomplexa, *Plasmodium* and *P. falciparum* Hsp70s were also mapped to the structures produced in this chapter (refer to Appendix I). These residues were too numerous and widely-spread across the protein to be analysed effectively. The models themselves may, however, be useful for future work, as they may grant insight into characteristics, specific to these proteins. Additionally they may be able to aid work regarding *in silico* drug design.

4. CONCLUSIONS AND FUTURE WORK

4.1 SEQUENCE FEATURES

This work has primarily involved identifying important sequence features within each of the four canonical PfHsp70 proteins and inspecting them at a structural level. In all PfHsp70s considered, putative J-domain contact residues and substrate contact residues and NEF contact residues were identified. As far as we are aware, this is the first time these residues have been characterised in PfHsp70-2, PfHsp70-3 and PfHsp70-x. This is also the first time putative Hsp110 contact residues have been identified in PfHsp70-1. To further the study the interaction of each PfHsp70 with other proteins requires the identification of these proteins in *P. falciparum* and modelling of these proteins in as they interact with their respective PfHsp70. The interaction of PfHsp70-1 with Hsp40 homologues (PfJs) has already been studied *in vitro*. It would be of great interest to simulate this binding *in silico*, as a reference for other potential PfHsp70 interactions with J-domain proteins. Another highly likely J-domain protein that has been identified is a *P. falciparum* Pam18 homologue, which would supposedly interact with PfHsp70-3. Modelling this interaction would grant insight into the effect of J-domain binding residue that appears to be missing from both PfHsp70-2 and PfHsp70-3 (an omission which is common to ER-Hsp70s and mtHsp70, respectively). This will indicate whether this residue is simply not important for J-domain protein binding in these types of Hsp70s or if the position of the next putative J-domain binding residue takes its place, as suggested by the structural analysis provided in Chapter 3. Unfortunately, there appears to be no structure in the PDB that represents the interaction of and Hsp70 with a J-domain, so modelling this interaction could prove difficult. Fortunately there are structures of different Hsp70s interacting with all three NEFs considered, so modelling this interaction should at least be possible. An additional consideration is the effect of the alpha-helical lid on the orientation of the substrate contact residues. Since the alpha helical lid is a highly variable region across Hsp70s, it might present itself as a viable drug target.

4.2 CONSERVED RESIDUES OF EACH PFHSP70

With regard to the study of the conserved features of the PfHsp70 proteins, a number of interesting discoveries were made. Firstly, the unusual sequence of PfHsp70-x, which has various features that are atypical of cytosolic Hsp70s and also has many residues conserved in

eukaryotes cytosolic Hsp70s, but not found in Hsp70s of *Plasmodium* or Apicomplexa. It is possible that PfHsp70-x represents a different type of Hsp70, with functions distinct from the other three types observed. The sequence information and model produced for this protein may be useful to supplement the study of PfHsp70-x *in vitro* and *in vivo*. The study of conserved sequence features in each type of Hsp70 has enabled the production of models with the conservation level of different residues mapped to the structures. These may be very useful for the purposes of designing antimalarial drugs, especially since these models are present in a variety of conformations. It may also be worth repeating this specific form of analysis, specifically using human Hsp70s, rather than a generalised eukaryotic reference. This would be more relevant to drug design against human malarial parasites. This is also not just limited to Hsp70s or malaria specifically and can be easily adjusted for use of other protein targets relevant to other diseases.

4.3 Modelling with Modified Templates

A novel approach to modelling has been suggested, which involves attempting to improve the quality of template structures before using them in the homology modelling process. It appears that this approach can be useful if the template in question is not a high quality structure. The structures used in this research were all of relatively high quality, so only very small improvements were ever noticed, if any. This technique is still being developed and future research efforts are required to determine its effectiveness when modelling proteins using relatively poor quality template structures. Additionally, this approach requires statistical analysis and a means to effectively determine whether or not the approach is viable and if so, over what range of template quality is it most effective. The current strategy regarding this will be to up the scale on which this is done, by acquiring hundreds of protein structures from the PDB. These can be divided into different groups or bins, based on their quality and used as templates for modelling. This approach will require finding a suitable target protein to be modelled using each template structure. If enough structures are used, it should allow the technique itself to be verified statistically. Also this will give an indication of whether or not the technique is more effective when templates of relatively poor quality are used and, if so, at what point the technique does become useful. Although this technique may prove to be of little value when modelling with high quality structures, it will hopefully prove useful for situations where a suitable high quality template cannot be identified for use in homology modelling.

5. REFERENCES

1. Acharya, P., Pallavi, R., Chandran, S., Chakravarti, H., Middha, S. & Acharya, J. (2009). A Glimpse into the Clinical Proteome of Human Malaria Parasites *Plasmodium Falciparum* and *Plasmodium Vivax*. *Proteomics. Clinical Applications*, *3*, 1314-25.

2. Aebi, A. C. S U. (2003). The Next Ice Age: Cryo-Electron Tomography of Intact Cells. *Trends in Cell Biology*, *13*(2), 107-10

3. Agarraberes, F. A. & Dice, J. F. (2001). A Molecular Chaperone Complex at the Lysosomal Membrane Is Required for Protein Translocation. *Journal of Cell Science*, *114*(13), 2491-9.

4. Ahmed, A., Whitford, P. C., Sanbonmatsu, K. Y. & Tama, F. (2011). Consensus Among Flexible Fitting Approaches Improves the Interpretation of Cryo-EM Data. *Journal of Structural Biology*, *177*(2), 561-70.

5. Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. (2007). The Design and Characterization of Two Proteins with 88% Sequence Identity but Different Structure and Function. *Proceedings of the National Academy of Sciences*, *104*(29), 11963-8.

6. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*, 403-10.

7. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z. & Miller, W. (1997). Gapped Blast and Psi-Blast: A New Generation of Protein Database Search Programs. *Nucleic Acids Research*, *25*(17), 3389-402.

8. Andersson, S. G. E., Karlberg, O., Canback, B. & Kurland, C. G. (2002). On the Origin of Mitochondria: A Genomics Perspective. *The Royal Society*, *358*, 165-79.

9. Anelli, T., Alessio, M., Mezghrani, A., Simmen, T., Talamo, F. & Bachi, A. (2002). ERp44, a Novel Endoplasmic Reticulum Folding Assistant of the Thioredoxin Family. *European Molecular Biology Organization Journal*, *21*(4), 835-44.

10. Arifuzzaman, M., Oshima, T., Nakade, S. & Mori, H. (2002). Characterization of HscC (Hsc62), Homologue of Hsp70 in Escherichia Coli: Over-Expression of HscC Modulates the Activity of House Keeping Sigma Factor Σ70. *Genes to Cells*, *7*, 553-66.

11. Aszodi, A. & Taylor, W. R. (1996). Homology Modelling by Distance Geometry. *Folding and Design*, *1*(5), 325-34.

12. Azem, A., Oppliger, W., Lustig, A., Jeno, P., Feifel, B. & Schatz, G. (1997). The Mitochondrial Hsp70 Chaperone System. *The Journal of Biological Chemistry*, *272*(33), 20901-6.

13. Baeten, L., Reumers, J., Tur, V., Stricher, F., Lenaerts, T. & Serrano, L. (2008). Reconstruction of Protein Backbones from the Brix Collection of Canonical Protein Fragments. *Public Library of Science Computational Biology*, *4*(5), 1-11

14. Bannister, L. H., Hopkins, J. M., Fowler, R. E., Krishna, S. & Mitchell, G. H. (2000). A Brief Illustrated Guide to the Ultrastructure of *Plasmodium Falciparum* Asexual Blood Stages. *Parasitology Today*, *16*(10), 427-33.

15. Bell, S. L., Chiang, A. N. & Brodsky, J. L. (2011). Expression of a Malarial Hsp70 Improves Defects in Chaperone-Dependent Activities in Ssa1 Mutant Yeast. *Public Library of Science*, *6*(5), 1-10.

16. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N. & Weissig, H. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235-42.

17. Bertelsen, E. B., Chang, L., Gestwicki, J. E. & Zuiderweg, E. R. P. (2009). Solution Conformation of Wild-Type *E. Coli* Hsp70 (DnaK) Chaperone Complexed with ADP and Substrate. *Proceedings of the National Academy of Sciences*, *106*(21), 8471-6.

18. Bimston, D., Song, J., Winchester, D., Takayama, S., Reed, J. C. & Morimoto, R. I. (1998). Bag-1, a Negative Regulator of Hsp70 Chaperone Activity, Uncouples Nucleotide Hydrolysis from Substrate Release. *The European Molecular Biology Organization Journal*, *17*(23), 6871-8.

19. Blond-Elguindi, S., Cwiria, S. E., Dower, W. J., Lipshutz, R. J., Sprang, S. R. & Sambrook, J. F. (1993). Affinity Panning of a Library of Peptides Displayed on Bacteriophages Reveals the Binding Specificity of BiP. *Cell*, *75*, 717-28.

20. Bohnert, M., Pfanner, N. & van der Laan, M. (2007). A Dynamic Machinery for Import of Mitochondrial Precursor Proteins. *Federation of the European Biochemical Societies Letters*, *581*, 2802-10.

21. Boniecki, M., Rotkiewicz, P., Skolnick, J. & Kolinski, A. (2003). Protein Fragment Reconstruction Using Various Modelling Techniques. *Journal of Computer-Aided Molecular Design*, *17*, 725-38.

22. Botha, M., Chiang, A. N., Needham, P. G., Stephens, L. L., Hoppe, H. C., Külzer, S., Przyborski, J. M., Lingelbach, K., Wipf, P., Brodsky, J. L., Shonhai, A. & Blatch, G. L. (2011). Plasmodium falciparum Encodes a Single Cytosolic Type I Hsp40 that Functionally Interacts with Hsp70 and is Upregulated by Heat Shock. *Cell Stress & Chaperones*, 16(4), 389-401.

23. Bowie, J. U., Luthy, R. & Eisenburg, D. (1991). A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science*, *253*(5016), 164-70.

24. Bradley, P., Misura, K. M. S. & Baker, D. (2005). Toward High-Resolution De Novo Structure Prediction for Small Proteins. *Science*, *309*, 1868-71.

25. Brehmer, D., Rüdiger, S., Gässler, C. S., Klostermeier, Dagmar, Packschies, L. & Reinstein, Jochen. (2001). Tuning of Chaperone Activity of Hsp70 Proteins by Modulation of Nucleotide Exchange. *Nature Structural Biology*, *8*(5), 427-432.

26. Buck, T. M., Wright, C. M. & Brodsky, J. L. (2007). The Activities and Function of Molecular Chaperones in the Endoplasmic Reticulum. *Seminars in Cell & Developmental Biology*, *18*, 751-61.

27. Chen, D-H., Baker, M. L., Hryc, C. F., DiMaio, F., Jakana, J., Wu, W., Dougherty, M., Haase-Pettingell, C., Schmid, M. F., Jiang, W., Baker, D., King, J. A. & Chiu, W. (2011). Structural Basis for Scaffolding-Mediated Assembly and Maturation of a dsDNA Virus. *Proceedings of the National Academy of Sciences, 108*(4), 1355-60

28. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J. & Higgins, D. G. (2003). Multiple Sequence Alignment with the Clustal Series of Programs. *Nucleic Acids Research*, *31*(13), 3497-500.

29. Chiang, A. N., Valderramos, J.-C., Balachandran, R., Chovatiya, R. J., Mead, B. P. & Schneider, C. (2009). Select Pyrimidinones Inhibit the Propagation of the Malarial Parasite, *Plasmodium Falciparum*. *Bioorganic & Medicinal Chemistry*, *17*, 1527-33. Elsevier Ltd.

30. Cowman, A. F. & Crabb, B. S. (2002). The *Plasmodium Falciparum* Genome - A Blueprint for Erythrocyte Invasion. *Science*, *298*(126), 126-8.

31. Desai, S. N., Agarwal, A. A. & Uplap, S. S. (2010). HSP: Evolved and Conserved Proteins, Structure and Sequence Studies. *International Journal of Bioinformatics Research*, *2*(2), 67-87.

32. D'silva, P. D., Schilke, B., Walter, W., Andrew, A. & Craig, E. A. (2003). J Protein Cochaperone of the Mitochondrial Inner Membrane Required for Protein Import into the Mitochondrial Matrix. *Proceedings of the National Academy of Sciences*, *100*(24), 13839-44.

33. di Luccio, E. & Koehl, P. (2011). A Quality Metric for Homology Modelling: The H-Factor. *Bioinformatics*, *12*, 48.

34. Easton, D. P., Kaneko, Y. & Subjeck, J. R. (2000). The Hsp110 and Grp170 Stress Proteins: Newly Recognized Relatives of the Hsp70s. *Cell Stress & Chaperones*, *5*(4), 276-90.

35. Edgar, R. C. (2004). Muscle: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research*, *32*(5), 1792-7.

36. Faou, P. & Hoogenraad, N. J. (2011). Tom34: A Cytosolic Cochaperone of the Hsp90/Hsp70 Protein Complex Involved in Mitochondrial Protein Import. *Molecular Cell Research*, *1823*(2), 348-57.

37. Fidelis, K., Stern, P. S., Bacon, D. & Moult, J. (1994). Comparison of Systematic Search and Database Methods for Constructing Segments of Protein Structure. *Protein Engineering*, *7*(8), 953-60.

38. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Graingerk, M. & Haynes, J. D. (2002). A Proteomic View of the *Plasmodium Falciparum* Life Cycle. *Nature*, *419*, 520-6.

39. Foster, M. P., Mcelroy, C. A. & Amero, C. D. (2007). Solution NMR of Large Molecules and Assemblies. *Biochemistry*, *46*(2), 331-40.

40. Foth, B. J., Ralph, S. A., Tonkin, C. J., Struck, N. S., Fraunholz, M. & Roos, D. S. (2003). Dissecting Apicoplast Targeting in the Malaria Parasite *Plasmodium Falciparum*. *Science*, *299*, 705-8.

41. Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M. & Hyman, R. W. (2002). Genome Sequence of the Human Malaria Parasite *Plasmodium Falciparum*. *Nature*, *419*, 498-511.

42. Gelinas, A. D., Toth, J., Bethoney, K. A., Stafford, W. F. & Harrison, C. J. (2004). Mutational Analysis of the Energetics of the GrpE·DnaK Binding Interface: Equilibrium Association Constants by Sedimentation Velocity Analytical Ultracentrifugation. *Journal of Molecular Biology*, *339*, 447-58.

43. Germot, A., Phillippe, H. & Le Guyader, H. (1996). Presence of a Mitochondrial-Type 70-kDa Heat Shock Protein in *Trichomonas Vaginalis* Suggests a Very Early Mitochondrial Endosymbiosis in Eukaryotes. *Proceedings of the National Academy of Sciences*, *93*, 14614-17.

44. Greenwood, B. M., Fidock, D. A., Kyle, D. E., Kappe, S. H. I., Alonso, P. L. & Collins, F. H. (2008). Review Series Malaria: Progress, Perils, and Prospects for Eradication. *The Journal of Clinical Investigation*, *118*(4), 1266-76.

45. Hall, T. A. (1999). Bioedit: A User-Friendly Biological Sequence Alignment Editor Program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, *41*, 95-8.

46. Hanssen, E., McMillan, P. J. & Tilley, L. (2010). Cellular Architecture of *Plasmodium Falciparum*-Infected Erythrocytes. *International Journal for Parasitology*, *40*(10), 1127-35.

47. Harrison, C. J., Hayer-Hartl, M., Liberto, M. D., Hartl, F. & Kuriyan, J. (1997). Crystal Structure of the Nucleotide Exchange Factor GrpE Bound to the ATPase Domain of the Molecular Chaperone DnaK. *Science*, *276*, 431-5.

48. Hartl, F. U. & Hayer-Hartl, M. (2002). Molecular Chaperones in the Cytosol: From Nascent Chain to Folded Protein. *Science*, *295*, 1852-8.

49. Hillisch, A., Pineda, L. F. & Hilgenfeld, R. (2004). Utility of Homology Models in the Drug Discovery Process. *Drug Discovery Today*, *9*(15), 659-69.

50. Hoogenraad, N. J., Ward, L. A. & Ryan, M. T. (2002). Import and Assembly of Proteins into Mitochondria of Mammalian Cells. *Biochimica Et Biophysica Acta*, *1592*, 97-105.

51. Illergård, K., Ardell, D. H. & Elofsson, Arne. (2009). Structure Is Three to Ten Times More Conserved Than Sequence - A Study of Structural Response in Protein Cores. *Proteins*, *77*, 499-508.

52. Inomata, K., Ohno, A., Tochio, H., Isogai, S., Tenno, T & Nakase, I. (2009). High-Resolution Multi-Dimensional NMR Spectroscopy of Proteins in Human Cells. *Nature*, *458*, 106-9.

53. Jiang, J., Prasad, K., Lafer, E. M. & Sousa, R. (2005). Structural Basis of Interdomain Communication in the Hsc70 Chaperone. *Molecular Cell*, *20*, 513-24.

54. Jung, G., Jones, G., Wegrzyn, R. D. & Masison, D. C. (2000). A Role for Cytosolic Hsp70 in Yeast [Psi+] Prion Propagation and [Psi +] As a Cellular Stress. *Genetics*, *156*, 559-70.

55. Kappes, B., Suetterlin, B. W., Hofer-Warbinek, R., Humar, R. & Franklin, R. M. (1993). Two Major Phosphoproteins of *Plasmodium Falciparum* Are Heat Shock Proteins. *Molecular and Biochemical Parasitology*, *59*, 83-94.

56. Katoh, K., Kuma, K. Toh, H. & Miyata, T. (2005). Mafft Version 5: Improvement in Accuracy of Multiple Sequence Alignment. *Nucleic Acids Research*, *33*(2), 511-8.

57. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). Mafft: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Research*, *30*(14), 3059-66.

58. Katoh, K. & Toh, H. (2008). Recent Developments in the Mafft Multiple Sequence Alignment Program. *Briefings in Bioinformatics*, *9*(4), 81-92.

59. Kim, S., Schilke, B., Craig, E. A. & Horwich, A. L. (1998). Folding *In Vivo* of a Newly Translated Yeast Cytosolic Enzyme Is Mediated by the Ssa Class of Cytosolic Yeast Hsp70 Proteins. *Proceedings of the National Academy of Sciences*, *95*, 12860-5.

60. Klepeis, J. L. & Floudas, C. A. (2003). Astro-Fold: A Combinatorial and Global Optimization Framework for Ab Initio Prediction of Three-Dimensional Structures of Proteins from the Amino Acid Sequence. *Biophysical Journal*, *85*(4), 2119-46.

61. Krieger, E., Nabuurs, S. B., & Vriend, G. (2003). Homology Modelling. *Structural Bioinformatics,* 44, ed. Bourne, P. & Weissig, H., John Wiley & Sons, Inc., Hoboken

62. Krishnamoorthy, B. & Tropsha, A. (2003). Development of a Four-Body Statistical Pseudo-Potential to Discriminate Native from Non-Native Protein Conformations. *Bioinformatics*, *19*(12), 1540-8.

63. Kumar, N., Koski, G., Harada, M., Aikawa, M. & Zheng, H. (1991). Induction and Localization of *Plasmodium Falciparum* Stress Proteins Related to the Heat Shock Protein 70 Family. *Molecular and Biochemical Parasitology*, *48*(1), 47-58.

64. Kumar, N., Syin, C., Carter, R., Quakyi, I. & Miller, L. H. (1988). *Plasmodium Falciparum* Gene Encoding a Protein Similar to the 78-Kda Rat Glucose-Regulated Stress Protein. *Proceedings of the National Academy of Sciences*, *85*, 6277-81.

65. Kun, J. & Muller-Hill, B. (1989). The Sequence of a Third Member of the Heat Shock Protein Family in *Plasmodium Falciparum*. *Nucleic Acids Research*, *17*(13), 8717-17.

66. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R, Mcgettigan, P. A. & Mcwilliam, H. (2007). Clustal W and Clustal X Version 2.0. *Bioinformatics*, *23*(21), 2947-8.

67. Laufen, T., Mayer, M. P., Beisel, C., Klostermeier, D., Mogk, A. & Reinstein, J. (1999). Mechanism of Regulation of Hsp70 Chaperones by DnaJ Cochaperones. *Proceedings of the National Academy of Sciences*, *96*(10), 5452-7.

68. Leslie, A. G. W. (2006). The Integration of Macromolecular Diffraction Data. *Acta Crystallographica*, *D62*, 48-57.

69. Levitt, M. (1983). Protein Folding by Restrained Energy Minimization and Molecular Dynamics. *Journal of Molecular Biology*, *170*, 723-64.

70. Levitt, M. & Lifson, S. (1969). Refinement of Protein Conformations Using a Macromolecular Energy Minimization Procedure. *Journal of Molecular Biology*, *46*, 269-79.

71. Lin, K., May, A. C. W. & Taylor, William R. (2002). Threading Using Neural Networks (TUNE): The Measure of Protein Sequence-Structure Compatibility. *Bioinformatics*, *18*(10), 1350-7.

72. Liu, Q. & Hendrickson, W. A. (2007). Insights into Hsp70 Chaperone Activity from a Crystal Structure of the Yeast Hsp110 Sse1. *Cell*, *131*, 106-20.

73. Liu, Q., Silva, P. D., Walter, W., Marszalek, J. & Craig, E. A. (2003). Regulated Cycling of Mitochondrial Hsp70 at the Protein Import Channel. *Science*, *300*, 139-41.

74. Luthy, R., Bowie, J. U. & Eisenburg, D. (1992). Assessment of Protein Models with Three-Dimensional Profiles. *Nature*, *356*, 83-5.

75. Matambo, T., Odunuga, O. O., Boshoff, A. & Blatch, G. L. (2004). Overproduction, Purification, and Characterization of the *Plasmodium Falciparum* Heat Shock Protein 70. *Protein Expression and Purification*, *33*, 214-22.

76. Matlack, K. E. S., Misselwitz, B., Plath, K. & Rapoport, T. A. (1999). BiP Acts As a Molecular Ratchet During Posttranslational Transport of Prepro-α Factor Across the ER Membrane. *Cell Press*, *97*, 553-64.

77. Mayer, M. P. & Bukau, B. (2005). Hsp70 Chaperones: Cellular Functions and Molecular Mechanism. *Cellular and Molecular Life Sciences*, *62*, 670-84.

78. Mayer, M. P., Schröder, H., Rüdiger, S., Paal, K., Laufen, T. & Bukau, B. (2000). Multistep Mechanism of Substrate Binding Determines Chaperone Activity of Hsp70. *Nature Structural Biology*, *7*(7), 586-93.

79. Melo, F. & Feytmans, E. (1998). Assessing Protein Structures with a Non-Local Atomic Interaction Energy. *Journal of Molecular Biology*, *277*, 1141-52.

80. Miao, B., Davis, J. E. & Craig, Elizabeth A. (1997). Mge1 Functions As a Nucleotide Release Factor for Ssc1, a Mitochondrial Hsp70 of Saccharomyces Cerevisiae. *Journal of Molecular Biology*, *265*, 541-52.

81. Miller, L. H., Baruch, D. I., Marsh, K. & Doumbo, O. K. (2002). The Pathogenic Basis of Malaria. *Nature*, *415*(6872), 673-9.

82. Mirny, L. A & Shakhnovich, E. I. (1998). Protein Structure Prediction by Threading. Why It Works and Why It Does Not. *Journal of Molecular Biology*, *283*, 507-26.

83. Misra, G. & Ramachandran, R. (2009). Hsp70-1 from *Plasmodium Falciparum*: Protein Stability, Domain Analysis and Chaperone Activity. *Biophysical Chemistry*, *142*, 55-64.

84. Miyata, T., Miyazawa, S. & Yasunaga, T. (1979). Two Types of Amino Acid Substitutions in Proteins. *Journal of Molecular Evolution*, *12*, 219-36.

85. Moro, F., Fernández-Sáiz, V. & Muga, A. (2004). The Lid Subdomain of DnaK is Required for the Stabilization of the Substrate-Binding Site. *The Journal of Biological Chemistry*, *279*(19), 19600-6.

86. Moseley, H. N. B. & Montelione, G. T. (1999). Automated Analysis of NMR Assignments and Structures for Proteins. *Current Opinion in Structural Biology*, 9, 635-42.

87. Munro, S. & Pelham, H. R. B. (1987). A C-Terminal Signal Prevents Secretion Luminal ER Proteins. *Cell*, *48*, 899-907.

88. Myers, E. W. & Miller, W. (1988). Optimal Alignments in Linear Space. *Computer Applications in the B*iosciences, 4(1), 1-13.

89. Ngom, A. (2006). Parallel Evolution Strategy on Grids for the Protein Threading Problem. *Journal of Parallel and Distributed Computing*, *66*, 1489-1502.

90. Nickell, S., Kofler, C., Leis, A. P. & Baumeister, W. (2006). A Visual Approach to Proteomics. *Nature*, *7*, 225-30.

91. Nicoll, W. S., Botha, M., McNamara, C., Schlange, M., Pesce, E-R. & Boshoff, A. (2007). Cytosolic and ER J-Domains of Mammalian and Parasitic Origin Can Functionally Interact with DnaK. *The International Journal of Biochemistry & Cell Biology*, *39*(4), 736-51.

92. Nyalwidhe, J. & Lingelbach, K. (2006). Proteases and Chaperones Are the Most Abundant Proteins in the Parasitophorous Vacuole of *Plasmodium Falciparum*-Infected Erythrocytes. *Proteomics*, *6*, 1563-73.

93. Olson, M. A., Feig, M. & Iii, C. L. B. (2007). Prediction of Protein Loop Conformations Using Multiscale Modelling Methods with Physical Energy Scoring Functions. *Journal of Computational Chemistry*, *29*(5), 820-31.

94. Pawlowski, M., Gajda, M. J., Matlak, R. & Bujnicki, J. M. (2008). MetaMQAP: A Meta-Server for the Quality Assessment of Protein Models. *Bioinformatics*, *9*, 403, 1-20.

95. Pei, J., Kim, B. H. & Grishin, N. V. (2008). PROMALS3D: A Tool for Multiple Protein Sequence and Structure Alignments. *Nucleic Acids Research*, *36*, 2295-300.

96. Pelham, H. R. B. (1986). Speculations on the Functions of the Major Heat Shock and Glucose-Regulated Proteins. *Cell*, *46*, 959-61.

97. Pelham, H. R. B. (1988). Evidence That Luminal ER Proteins Are Sorted from Secreted Proteins in a Post-ER Compartment. *European Molecular Biology Organization Journal*, *7*(4), 913-8.

98. Pellecchia, M., Montgomery, D. L., Stevens, S. Y., Kooi, C. W. V., Feng, H.-ping, Gierasch, L. M. & Zuiderweg, E. R. P. (2000). Structural Insights into Substrate Binding by the Molecular Chaperone DnaK. *Nature Structural Biology*, 7(4), 298-303.

99. Pesce, E-R., Acharya, P., Tatu, U., Nicoll, W. S., Shonhai, A. & Hoppe, H. C. (2008). The *Plasmodium Falciparum* Heat Shock Protein 40, Pfj4, Associates with Heat Shock Protein 70 and Shows Similar Heat Induction and Localisation Patterns. *The International Journal of Biochemistry & Cell Biology*, *40*, 2914-26.

100. Peterson, M. G., Crewther, P. E., Thompson, J. K., Corcoran, L. M., Coppel, R. L. & Brown, G. V. (1988). A Second Antigenic Heat Shock Protein of *Plasmodium Falciparum. DNA Research*, *7*(2), 71-8.

101. Polier, S., Dragovic, Z., Hartl, F. U. & Bracher, A. (2008). Structural Basis for the Cooperation of Hsp70 and Hsp110 Chaperones in Protein Folding. *Cell*, *133*, 1068-79.

102. Ponder, J. W. & Richards, F. M. (1987). Tertiary Templates for Proteins. Use of Packing Criteria in the Enumeration of Allowed Sequences for Different Structural Classes. *Journal of Molecular Biology*, *193*, 775-91.

103. Radermacher, M., Wagenknecht, T., Grassucci, R., Frank, J., Inui, M. & Chadwick, C. (1992). Cryo-EM of the Native Structure of the Calcium Release Channel/Ryanodine Receptor from Sarcoplasmic Reticulum. *Biophysical Journal*, *61*, 936-40.

104. Ramya, T. N. C., Karmodiya, K., Surolia, A. & Surolia, N. (2007). 15-Deoxyspergualin Primarily Targets the Trafficking of Apicoplast Proteins in *Plasmodium Falciparum*. *Journal of Biological Chemistry*, *282*(9), 6388-97.

105. Ramya, T. N. C., Surolia, N. & Surolia, A. (2006). 15-Deoxyspergualin Modulates *Plasmodium Falciparum* Heat Shock Protein Function. *Biochemical and Biophysical Research Communications*, *348*, 585-92.

106. Ran, Q., Wadhwa, R., Kawai, R., Kaul, S. C., Sifers, R. N. & Bick, R. J. (2000). Extramitochondrial Localization of Mortalin/MtHsp70/Pbp74/Grp75. *Biochemical and Biophysical Research Communications*, *275*, 174-9.

107. Roessler, C. G., Hall, B. M., Anderson, W. J., Ingram, W. M., Roberts, S. A. & Montfort, W. R. (2008). Transitive Homology-Guided Structural Studies Lead to Discovery of Cro Proteins with 40% Sequence Identity but Different Folds. *Proceedings of the National Academy of Sciences*, *105*(7), 2343-8.

108. Rost, B. (1999) Twilight Zone of Protein Sequence Alignments. *Protein Engineering*. *12*(2), 85-94

109. Rudiger, S., Germeroth, L., Schneider-Mergener, J. & Bukau, Bernd. (1997). Substrate Specificity of the DnaK Chaperone Determined by Screening Cellulose-Bound Peptide Libraries. *The European Molecular Biology Organization Journal*, *16*(7), 1501-7.

110. Saitou, N. & Nei, M. (1987). The Neighbour-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, *4*(4), 406-25.

111. Sakuragi, S. Liu, Q. & Craig, E. (1999). Interaction between the Nucleotide Exchange Factor Mge1 and the Mitochondrial Hsp70 Ssc1. *Journal of Biological Chemistry*, *274*(16), 11275-82.

112. Sali, A. (2009) MODELLER: A Program for Protein Structure Modelling. *University of California*

113. Sali, A. & Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Biological Chemistry*, *234*, 779-815.

114. Sali, A., Ovedngton, J. P., Johnson, M. S. & Blundell, T. L. (1990). From Comparisons of Protein Sequences and Structures to Protein Modelling and Design. *Trends in Biochemical Sciences*, *15*, 235-40.

115. Sargeant, T. J., Marti, M., Caler, E., Carlton, J. M., Simpson, K. & Speed, T. P. (2006). Lineage-Specific Expansion of Proteins Exported to Erythrocytes in Malaria Parasites. *Genome Biology*, *7*(2), R12.1-23.

116. Schlecht, R., Erbse, A. H., Bukau, B. & Mayer, M. P. (2011). Mechanics of Hsp70 Chaperones Enables Differential Interaction with Client Proteins. *Nature Structural & Molecular Biology*, *18*(3), 345-51.

117. Schuermann, J. P., Jiang, J., Cuellar, J., Llorca, O., Wang, L. & Gimenez, L. E. (2008). Structure of the Hsp110:Hsc70 Nucleotide Exchange Machine. *Molecular Cell*, *31*(2), 232-43.

118. Shen, M.-Y. & Sali, A. (2006). Statistical Potential for Assessment and Prediction of Protein Structures. *Protein Science*, *15*, 2507-24.

119. Shomura, Y., Dragovic, Z., Chang, H-C., Tzvetkov, N., Young, J. C. & Brodsky, J. L. (2002). Regulation of Hsp70 Function by Hspbp1: Structural Analysis Reveals an Alternate Mechanism for Hsp70 Nucleotide Exchange. *Molecular and Cellular Biology*, *17*, 367-79.

120. Shonhai, A., Boshoff, A. & Blatch, G. L. (2005). *Plasmodium Falciparum* Heat Shock Protein 70 Is Able to Suppress the Thermosensitivity of an *Escherichia Coli* DnaK Mutant Strain. *Molecular Genetics and Genomics*, *274*, 70-8.

121. Shonhai, A., Boshoff, A. & Blatch, G. L. (2007). The Structural and Functional Diversity of Hsp70 Proteins from *Plasmodium Falciparum*. *Protein Science*, *16*, 1803-18.

122. Silberg, J. J., Hoff, K. G., Vickery, L. E., Silberg, J. J., Hoff, K. G. & Vickery, L. E. (1998). The Hsc66-Hsc20 Chaperone System in *Escherichia Coli*: Chaperone Activity and Interactions with the DnaK-DnaJ-GrpE System. *Journal of Bacteriology*, *180*(24), 6617-24.

123. Singh, B., Sung, L. K., Matusop, A., Radhakrishnan, A., Shamsul, S. S. G. & Cox-Singh, J. (2004). A Large Focus of Naturally Acquired *Plasmodium Knowlesi* Infections in Human Beings. *The Lancet*, *363*, 1017-24.

124. Sippl, M. J. (1990). Calculation of Conformational Ensembles from Potentials of Mean Force. *Journal of Molecular Biology*, *213*, 859-83.

125. Sippl, M. J. (1993). Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins: Structure, Function, and Genetics*, *17*, 355-62.

126. Slapeta, J. & Keithly, J. S. (2004). Cryptosporidium Parvum Mitochondrial-Type Hsp70 Targets Homologous and Heterologous Mitochondria. *Eukaryotic Cell*, *3*(2), 483-94.

127. Smyth, M. S. & Martin, J. H. J. (2000). X-Ray Crystallography. *Journal of Clinical Pathology: Molecular Pathology*, *53*, 8-14.

128. Solomon, I. (1955). Relaxation Processes in a System of Two Spins. *Physical Review*, *99(2)*, 559-65

129. Sondermann, H., Scheuffler, C., Schneider, C., Hohfeld, J., Hartl, F. U. & Moarefi, I. (2001). Structure of a Bag/Hsc70 Complex: Convergent Functional Evolution of Hsp70 Nucleotide Exchange Factors. *Science*, *291*, 1553-7.

130. Striepen, B., Jordan, C. N., Reiff, S. & van Dooren, G. G. (2007). Building the Perfect Parasite: Cell Division in Apicomplexa. *Public Library of Science*, *3*(6), 691-8.

131. Söding, J., Biegert, A. & Lupas, A. N. (2005). The HHpred Interactive Server for Protein Homology Detection and Structure Prediction. *Nucleic Acids Research*, *33*(Web Server Issue), W244-8.

132. Takayama, K., Nakase, I., Michiue, H., Takeuchi, T., Tomizawa, K. & Matsui, H. (2009). Enhanced Intracellular Delivery Using Arginine-Rich Peptides by the Addition of Penetration Accelerating Sequences (Pas). *Journal of Controlled Release*, *138*, 128-33.

133. Tardieux, I., Baines, I., Mossakowska, M. & Ward, G. E. (1998). Actin-Binding Proteins of Invasive Malaria Parasites and the Regulation of Actin Polymerization by a Complex of 32/34-kDa Proteins Associated with Heat Shock Protein 70kDa. *Molecular and Biochemical Parasitology*, 93(2), 295-308.

134. Thompson, Julie D, Higgins, Desmond G & Gibson, Toby J. (1994). Clustal W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research*, *22*(22), 4673-80.

135. Tina, K. G., Bhadra, R. & Srinivasan, N. (2007). PIC: Protein Interactions Calculator. *Nucleic Acids Research*, *35*, 473-6.

136. Topf, M. & Sali, A. (2005). Combining Electron Microscopy and Comparative Protein Structure Modelling. *Current Opinion in Structural Biology*, *15*, 578-85.

137. Trott, O. & Olson, A. J. (2009). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry*, *31*(2), 455-61.

138. Tsai, J. & Douglas, M. G. (1996). A Conserved HPD Sequence of the J-Domain Is Necessary for Ydj1 Stimulation of Hsp70 ATPase Activity at a Site Distinct from Substrate Binding. *The Journal of Biological Chemistry*, *271*(16), 9347-54.

139. van Dooren, G. G., Marti, M., Tonkin, C. J., Stimmler, L. M., Cowman, A. F. & Mcfadden, G. I. (2005). Development of the Endoplasmic Reticulum, Mitochondrion and Apicoplast During the Asexual Life Cycle of *Plasmodium Falciparum*. *Molecular Microbiology*, *57*(2), 405-19.

140. Venclovas, C. (2003). Comparative Modelling in CASP5: Progress Is Evident, but Alignment Errors Remain a Significant Hindrance. *Proteins*, *53*, 380-8.

141. Vlijmen, H. W. Van & Karplus, M. (1997). PDB-Based Protein Loop Prediction: Parameters for Selection and Methods for Optimization. *Journal of Molecular Biology*, *267*, 975-1001.

142. Vogel, J. P., Misra, L. M. & Rose, M. D. (1990). Loss of BiP/Grp78 Function Blocks Translocation of Secretory Proteins in Yeast. *The Journal of Cell Biology*, *110*, 1885-95.

143. Wadhwa, R., Taira, K. & Kaul, S. C. (2002). An Hsp70 Family Chaperone, Mortalin/MtHsp70/Pbp74/Grp75: What, When, and Where? *Cell Stress & Chaperones*, *7*(3), 309-16.

144. Waller, R. F. & Mcfadden, G. I. (2004). The Apicoplast: A Review of the Derived Plastid of Apicomplexan Parasites. *Current Issues of Molecular Biology*, *7*, 57-80.

145. Wallner, B. & Elofsson, Arne. (2006). Identification of Correct Regions in Protein Models Using Structural, Alignment, and Consensus Information. *Protein Science*, *15*, 900-13.

146. Wiederstein, M. & Sippl, M. J. (2007). ProSA-Web: Interactive Web Service for the Recognition of Errors in Three-Dimensional Structures of Proteins. *Nucleic Acids Research*, *35*(Web Server Issue), W407-10.

147. Wüthrich, K. (2001). The Way to NMR Structures of Proteins. *Nature Structural Biology*, *8*(11), 923-5.

148. Xiang, Z & Honig, B. (2001). Extending the Accuracy Limits of Prediction for Side-Chain Conformations. *Journal of Molecular Biology*, *311*, 421-30.

149. Xiang, Z. (2006). Advances in Homology Protein Structure Modelling. *Current Protein & Peptide Science*, *7*(3), 217-27.

150. Yoshimune, K., Yoshimura, T., Nakayama, T., Nishino, T. & Esaki, N. (2002). Hsc62, Hsc56, and GrpE, the Third Hsp70 Chaperone System of Escherichia Coli. *Biochemical and Biophysical Research Communications*, *293*, 1389-95.

151. Young, J. C., Moarefi, I. & Hartl, F. U. (2001). Hsp90: A Specialized but Essential Protein-Folding Tool. *The Journal of Cell Biology*, *154*(2), 267-73.

152. Zelma, A., Venclovas, C., Moult, J. & Fidelis, K. (2001). Processing and Evaluation of Predictions in CASP4. *Proteins*, *5*, 13-21.

153. Zhang, X. P., Elofsson, A, Andreu, D. & Glaser, E. (1999). Interaction of Mitochondrial Presequences with DnaK and Mitochondrial Hsp70. *Journal of Molecular Biology*, *288*, 177-90.

154. Żmijewski, M. A., Skórko-Glonek, J., Tanfani, F., Banecki, B., Kotlarz, A. & Macario, A. J. L. (2007). Structural Basis of the Interspecies Interaction between the Chaperone DnaK (Hsp70) and the Co-Chaperone GrpE of Archaea and Bacteria. *Acta Biochimica Polonica*, *54*, 245-52.