In-silico analysis of *Plasmodium falciparum* Hop protein and its interactions with Hsp70 and Hsp90

A mini-thesis submitted in partial 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

Crystal-Leigh Clitheroe

February 2013

Acknowledgements

I wish to acknowledge all the people and institutions without whom I would not have been able to finish this work.

Most importantly, my sincerest thanks to my supervisor, Dr Özlem Tastan Bishop, for pushing me, for her patience, criticism, academic expertise, advice, encouragement and for her faith in me. I would also like to thank all the lecturers and academics associated with the RUBi Bioinformatics Master's course for their instruction and advice this year. Thanks also to Schmid et al (2012) for kindly allowing us to use their unpublished structure of ScHopTPR2 in complex with ScHsp90, referred to as "SchmidCYS" in this work).

My endless thanks to my mother, Theresa Clitheroe, to whom I attribute all the *best* parts of me, for all her sacrifices for me and for always believing in me. Also, thanks to my brother, Hillel Clitheroe, of whom I am very proud, for his love and for constantly reminding me not to become an old woman.

Especially, I wish to thank Matthew and Candice Coombes for their love, support and encouragement, as well asfor their criticism, advice and assistance with editing.

Thanks to Lyndall Perriera, for her kindness, friendship and support, as well as advice on phylogenetic analyses.

I wish to thank all my colleagues in RUBi, particularly Candice Ryan and Rowan Hatherley, for their camaraderie, feedback, ideas and academic advice on all things.

I would also like to thank the Rhodes University Post-graduate Financial Aid Office staff, Liezel Knott and John Gillam, for their assistance and advice, as well as the Rhodes University Sandisa Imbewu Fund for funding my studies with a Master's scholarship in 2012. Thanks also to the Department of Science and Technology (DST) for several Student Travel Fellowships to fund trips Stellenbosch and Johannesburg to attend EnsEMBL and Galaxy workshops, as well as to present this work at the National SAGS and Bioinformatics Conference in 2012.

Dedication

This thesis is lovingly dedicated to the memory of my grandparents,

Janine and Dieter Apfelthaler

Abstract

A lessor understood co-chaperone, the Hsp70/Hsp90 organising protein (Hop), has been found to play an important role in modulating the activity and co-interaction of two essential chaperones; Hsp90 and Hsp70. The best understood aspects of Hop so far indicate that residues in the concave surfaces of the three tetratricopeptide repeat (TPR) domains in the protein bind selectively to the C-terminal motifs of Hsp70 and Hsp90. Recent research suggests that *P. falciparum* Hop (PfHop), PfHsp90 and PfHsp70 do interact and form complex in the *P. falciparum* trophozooite and are overexpressed in this infective stage. However, there has been almost no computational research on malarial Hop protein in complex with other malarial Hsps.The current work has focussed on several aspects of the *in-silico* characterisation of PfHop, including an in-depth multiple sequence alignment and phylogenetic analysis of the protein; which showed that Hop is very well conserved across a wide range of available phyla (four Kingdoms, 60 species). Homology modelling was employed to predict several protein structures for these interactions in *P. falciparum*, as well as predict structures of the relevant TPR domains of Human Hop (HsHop) in complex with its own Hsp90 and Hsp70 C-terminal peptide partners for comparison. Protein complex interaction analyses indicate that concave TPR sites bound to the C-terminal motifs of partner proteins are very similar in both species, due to the excellent conservation of the TPR domain's "double carboxylate binding clamp". Motif analysis was combined with phylogenetic trees and structure mapping in novel ways to attain more information on the evolutionary conservation of important structural and functional sites on Hop. Alternative sites of interaction between Hop TPR2 and Hsp90's M and C domains are distinctly less well conserved between the two species, but still important to complex formation, making this a likely interaction site for selective drug targeting. Binding and interaction energies for all modelled complexes have been calculated; indicating that all HsHop TPR domains have higher affinities for their respective C-terminal partners than do their *P. falciparum* counterparts. An alternate motif corresponding to the C-terminal motif of PfHsp70-x (exported to the infected erythrocyte cytosol) in complex with both human and malarial TPR1 and TPR2B domains was analysed, and these studies suggest that the human TPR domains have a higher affinity for this motif than do the respective PfHop TPR domains. This may indicate potential for a cross species protein interaction to take place, as PfHop is not transported to the human erythrocyte cytosol.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction and Literature Review

1.1 Malaria

Malaria is a febrile illness caused by parasites of the apicomplexa phylum (Alberts et al., 2002), primarily by five species of parasites of the genus P*lasmodium* that affect humans (*P. falciparum, P. vivax, P. chabaudi, P. yoeleii* (Alberts et al., 2002) and the more recently discovered *P. knowlesi* (WHO: Global Malaria Programme, 2011). Malaria contracted from *P. falciparum* is the most deadly (cerebral infection), and it predominates in Africa. *P. vivax* is less dangerous but more widespread and the other three species are found much less frequently (Alberts et al., 2002). The malaria-causing protist is transmitted to humans (primary host) exclusively through the bite of malaria-infected, pregnant female mosquitoes (vector hosts) of any of 60 *Anopheles* species (Alberts et al., 2002).

In 2010, there were approximately 216 million reported cases of malaria and an estimated 655 000 deaths (WHO: Global Malaria Programme, 2011). The worldwide death rates due malaria have been reported to have decreased by approximately 25% since 2000, and by 33% in the WHO African Region, however it was estimated that 3.3 billion people were at risk of contracting malaria at any one time in 2010. In that year, with an estimated 81% of cases and 91% of all malaria related deaths in all geographic regions, the WHO African Region populations were at the highest risk of acquiring malaria. Pregnant women and children younger than five years have been the most severely affected; the disease is responsible for approximately 22% of all childhood deaths (WHO: Global Malaria Programme, 2011).

1.2 Parasite Life Cycle

The malaria-causing protist is transmitted to humans (primary host) exclusively through the bite of malaria-infected, pregnant female mosquitoes (vector hosts) of any of 60 anopheles species (Alberts et al., 2002). *P. falciparum* exists in at least 8 distinct forms, requiring both the human and mosquito hosts to complete its sexual cycle (Alberts et al., 2002). Gametes are formed in the blood stream of the human host but can only fuse to form a zygote in the gut of the anopheles mosquito. The greatest challenges facing the

parasite during human infection are overcoming change in temperature and pH from mosquito midgut to the human bloodstream (Alberts et al., 2002).

The mosquito midgut environment, which is approximately ambient and pH 8.5-9.5 respectively, is so basic due to the increase in pH observed during digestion of a blood meal, resulting in conversion of carbon dioxide to bicarbonate via carbonic anhydrase activity (Corena et al., 2005). This activity is essential to parasite development as it helps to induce gametogenesis of *Plasmodium* parasites (del Pilar Corena et al., 2005).On the other hand, the mature red blood cell is a characteristically nutrient poor environment, but the modification of the parasite-infected cell allows the red-blood cells to become permeable to influx of small molecules (via diffusion), making this environment consistent with the surrounding blood plasma, (37°C and pH 7.34).

In the human body, the malaria parasite migrates to the liver to invade hepatocytes, where it develops through the 'ring' and trophozoite stages, then divides to form an average of 20 daughter cells in the schizont stage, which in turn go on to infect red blood cells (Tilley, Dixon, & Kirk, 2011).This makes it unique in contrast to a number of other intracellular parasites which invade the nucleated, metabolically active cells that allow easy access to cellular ingredients provided by the host cell (Charpian et al., 2008).

Thus the parasite, trapped within the red blood cell must find its own way to access nutrients from the extracellular environment (i.e. the blood plasma). To deal with these constraints, the parasite has developed cryptic methods of transporting a number of select proteins (critical for mediating functions such as nutrient acquisition, cytoadherance and evasion of the human immune system) to the host erythrocyte cytoplasm and plasmalemma. It does this, for reasons that are still not completely understood, by inducing the formation of a parasitophorous vacuole (see Figure 1.1), creating a compartment within the red blood cell, enclosed by the parasitophorous vacuole membrane (PVM) (Charpian et al., 2008).

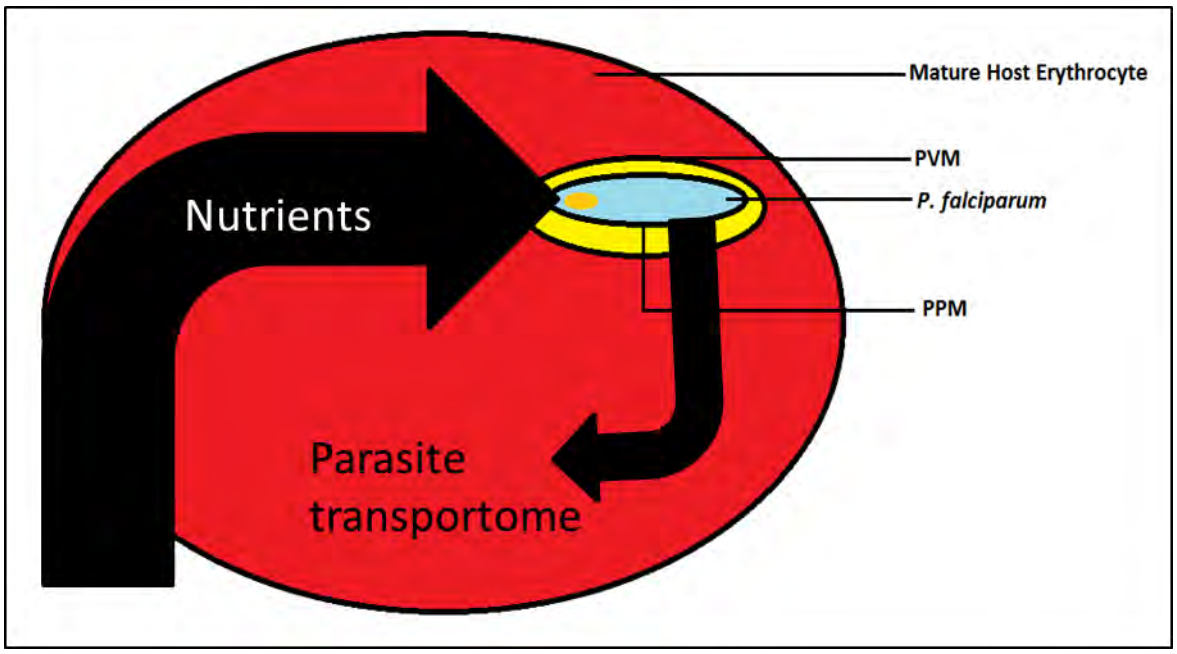


Figure 1.1: The parasitic vacuole system of malaria in a red blood cell. PPM and PVM are the parasite plasma membrane and parasitophorous vacuole membrane, respectively.

In order to acquire nutrients for parasitic growth, proteins must cross both the parasite plasma membrane (PPM) and the PVM. It has been indicated that transported parasite proteins contain short peptide sequences, such as the PEXEL (Plasmodium export element) motif, that seems to direct traffic of many soluble and membrane proteins to the infected erythrocyte by means as yet to be elucidated (Charpian et al., 2008). However, there are other proteins that do not possess these PEXEL (or similar) motifs that have been found to be transported to the host cytosol. An example of this is a *P. falciparum* heat-shock protein 70 variant (PfHsp70-x), which was characterised *in-silico,* found not to possess this PEXEL motif (Hatherley, 2012) and its transport to the host cytosol was recently determined experimentally (Kulzer et al, 2012).

1.3 Malaria Treatment and Drug Resistance

The biggest problem for traditional treatment of malaria with chloroquine and other more recently used drug combinations is the development of drug resistance in the parasite. Resistant *P. falciparum* strains have been found to overexpress an ABC transporter

membrane protein in the presence of chloroqine, which actively pumps the drug out of the parasite cell (Alberts et al., 2002). As such, the search for new drug targets is ongoing. Heat shock proteins (Hsps) seem to be a good choice, as they assist the parasite in coping during the transferral from mosquito midgut environment to the human bloodstream and may even play a role in helping the parasite to develop resistance (Gitau et al, 2012). Most Hsps are well conserved across taxa, so human and *P. falciparum* proteins are very similar. This may cause problems for selective drug targeting but also safeguards against developing resistance, as the genes expressing Hsps are less likely to mutate.

1.4 Heat Shock Proteins

The malaria parasites' Hsps are responsible for several of its key biological mechanisms. Most important is the thermo-protective, primary chaperone function, assisting the parasite to survive the transition from mosquito midgut to the human bloodstream, and then adaptation to two different cellular environments; mature erythrocytes and hepatocytes (Shonhai, 2010). Correct transmembrane export unfolding and refolding (Charpian et al., 2008) is critical to the parasites pathogenicity and allows it to survive the relatively nutrient and protein poor environment of the red blood cell (Charpian et al., 2008; Shonhai, 2010). This allows facilitating parasite interorganelle protein trafficking and regulation of parasite growth, infectivity and pathogenicity (Shonhai, 2010).

1.5 Hsp70/Hsp90 Organising Protein (Hop)

Through several years of research and analyses inyeast, the Hsp90 complex, including Sti1/Hop,is classified as a stress-inducible chaperone complex (Albanèse et al., 2006). Hop was first mentioned in 1986 and 1989 (Giordano et al., 1989) where a 62kD protein was immunoprecipitated with E1A protein and other heat shock proteins in adenovirus infected cells. It was isolated and characterised in yeast as Sti1 – Stress Inducible protein 1 (Nicolet & Craig, 1989). It was first identified in human as IEF SSP 3521, as it was over expressed after simian 40 virus transformation (Honore et al., 1992). It has also been described as p60 in aves (Smith et al., 1993). It has since been predicted from several species' genomes (Odunuga, Longshaw, & Blatch, 2004) and was recently characterised in *P. Falciparum* (Gitau et al., 2012).

1.5.1 Intracellular Localization of Hop

The distribution of Hop in the cell, as well as in the body seems to be extremely widespread. It has been found to have a primarily cytoplasmic distribution in mouse cells (Lässle et al., 1997). However, approximately 6% of the cellular fraction is thought to be associated with the Golgi apparatus and vesicles (Honore et al., 1992), cell surface (Martins et al., 1997; Zanata et al., 2002) and within the membrane fraction (Mehrpour et al.,2010; Sakadu et al., 2005). Certain conditions including G1/S arrest et al., 2004) and heat-shock or treatment with leptomycin B promote nuclear localization in mouse, and two nuclear localisation signals have been identified in the HsHop protein sequence (Daniel et al., 2008). Both mouse Hop and HsHop have also been reported to be recruited to stress granules along with several other cochaperones (Lapointe, Lasko, & Hobman, 2009).

1.5.2 Extracellular Localization of Hop

There is a large amount of evidence to suggest that Hop is externally localised, primarily in the mammalian brain tissue. However, there is also some research being done on Hop secretion by other tissue culture cells in mice (Eustace & Jay, 2004; Lima et al., 2007) and ovarian cancer cells in humans (Wang et al., 2010). Hsp90 complex secreted with co-chaperones p23, Hop, Hsp70 and Hsp40 increases the ATP-independent activation of matrix metalloproteinase 2 (MMP-2) in a breast cancer cell-line (Sims, McCready, & Jay, 2011). Proteins of the MMP family are involved in the breakdown of extracellular matrix important for normal physiological processes, such as embryonic development, reproduction, and tissue remodeling, as well as in disease processes, such as arthritis and metastasis (Devarajan et al., 1992).

Extracellular Hop may be able to differentiate between normal glial cells and glioma (brain or spinal tumour cells), as it stimulates the proliferation of only glioma through activation of the MAPK (Americo et al., 2007) and PI3K pathways in mice and rats (Erlich et al., 2007). Intra-hippocampal infusion of antibodies to Hop have been found to impair memory in rats, hence the interest in this protein for Alzheimers research (Coitinho et al., 2007). In mice, it has been found that there is a prion protein (PrP^c) interacting peptide of Hop secreted by astrocytes (Lima et al., 2007) that confers neuroprotection (Romano et al., 2007; Zanata et al., 2002), dependent on interaction with and activation of certain kinases (Lopes et al., 2005). PrP^c-dependent stimulation of translation by Hop has been found to be mediated by mTOR signalling, which is in turn activated by several cell processes of

interest to disease related research; tumor formation and angiogenesis, insulin resistance, adipogenesis and T-lymphocyte activation (Roffé et al., 2010). Hop also modulates activity of α7 nicotinic acetylcholine receptor for which PrP^c may act as receptor or co-receptor in mice (Roffé et al., 2010). Additionally, Hop:PrP^ccomplex formation may play a role in neurosphere formation in mice, and thus promote neural stemness (Weiss & Dos Santos, 2009).

1.5.3 Hsp90 Chaperone Suite and the Role of Hop

In an overview of two models for Hsp chaperone complex functioning recently put forward by Southworth and Agard (2011) and Schmid et al. (2012), yeast Hop (ScHop) first binds to an Hsp90 dimer, stabilizing the Hsp90 client-loading conformation by inhibiting ATP binding. Hop consists of a TPR1, DP1, TPR2A, TPR2B and DP2 region (see Figure 1.2). The TPR1–DP1 fragment is connected by a long flexible linker region to the rigid TPR2A–TPR2B block (Schmid et al., 2012). The DP2 domain is linked to TPR2B by another short linker region.

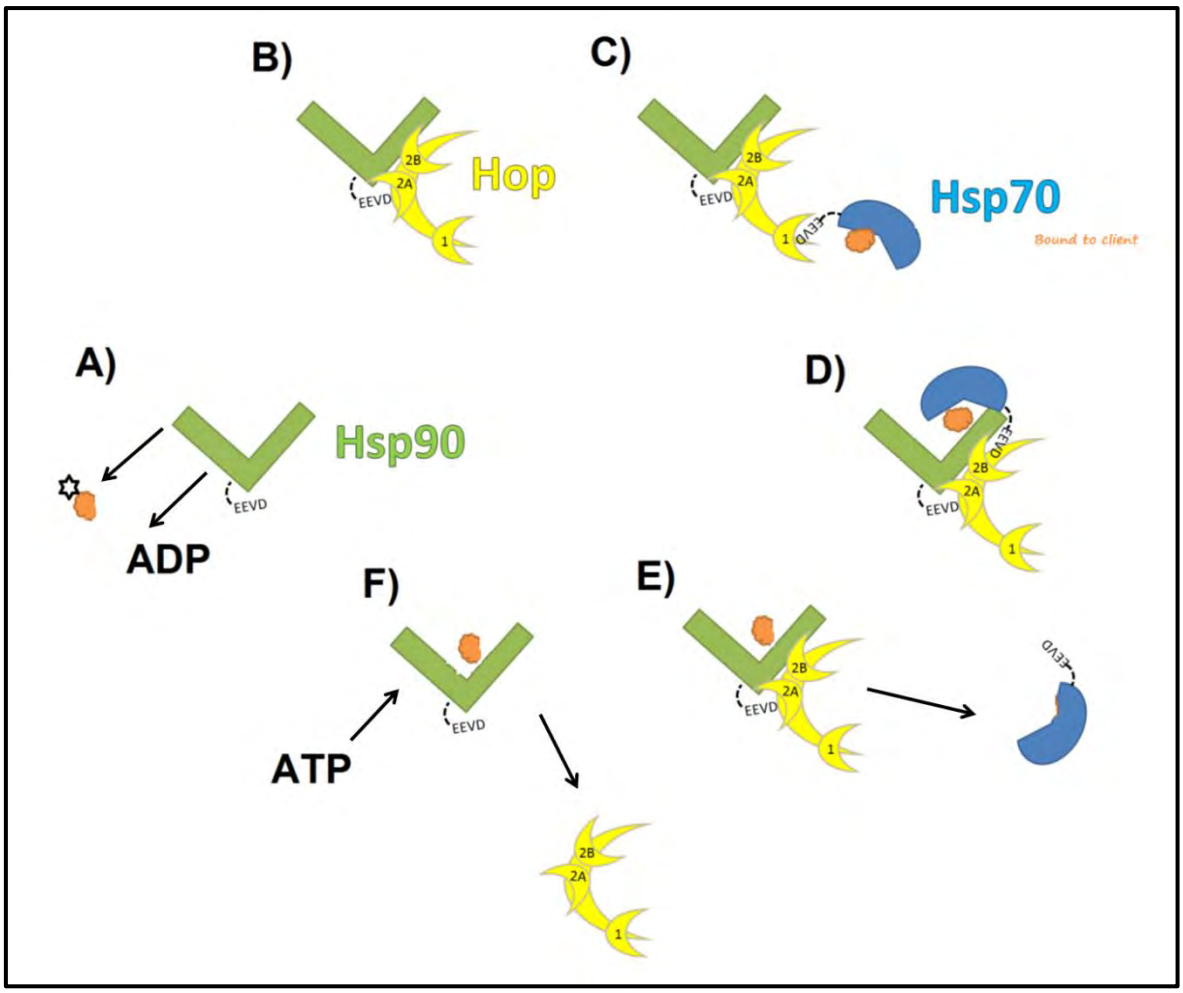


Figure 1.2: Simplistic overview of the Hsp90:Hop:Hsp70:client chaperone suite formation, adapted from Southworth & Agard (2012).A) Hsp90 homodimer in relaxed, open conformation. B) One of the Hsp90 C-terminal EEVD peptides in the Hsp90 homodimer bind TPR2A while the Hsp90-M domain interacts with TPR2B, inhibiting Hsp90 ATPase activity. C) Hsp70:client complex then binds to TPR1, via the C-terminal EEVD motif on Hsp70. D) Hsp70 and client are transferred to TPR2B. E) Client is transferred to the Hsp90's hydrophobic, homodimer cleft and Hsp70 releases. F) Hop releases, allowing ATP to bind and processing of the client protein by Hsp90. The cycle is completed after ATP hydrolysis and release of ADP and the activated client occurs, and Hsp90 returns to the relaxed open state in A.

In the Hsp90:Hop complex, one of the Hsp90 C-terminals in the Hsp90 homodimer binds the concave surface of TPR2A while the Hsp90-M domain interacts with the covex surface of TPR2B (see Figure 1.2 A and B), leading to the inhibition of the Hsp90 ATPase (Southworth & Agard, 2011a). The Hsp70:client complex then binds, owing to higher affinity, Hsp70 C-terminal is initially bound to TPR1 and it is thought that DP1 may stabilize the bound client (Schmid et al., 2012), interacting with Hop and Hsp90 (see Figure 1.2 C). Subsequently, Hsp70 and client are transferred to TPR2B–DP2 (Figure 1.2 D) which facilitates the release of the client protein to hydrophobic residues in the Hsp90

interdimer cleft (Southworth and Agard, 2011). From this position, the client is then transferred to Hsp90. Once Hop and Hsp70 are released, and ATP can bind Hsp90, the Hsp90 N-Terminal domains dimerize, forming the closed state represented in Figure 1.2 F. Previous studies on this mechanism of the ATPase inhibition suggest that the closed Hsp90 does not entirely enclose its client proteins but provides a bipartite binding surface whose formation and inhibition are coupled to the chaperone ATPase cycle (Ali et al., 2006).The cycle is completed after ATP hydrolysis and release of ADP and the activated client, and Hsp90 returns to a relaxed open state in Figure 1.2 A (Southworth &Agard, 2011).

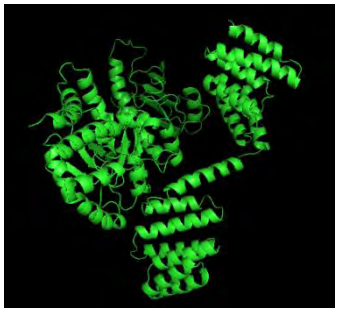
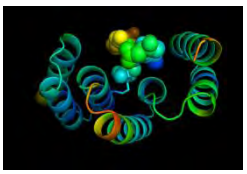
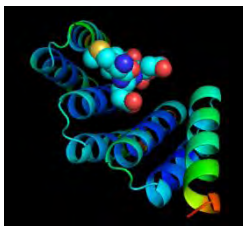
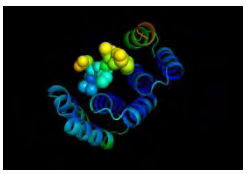
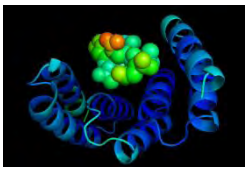
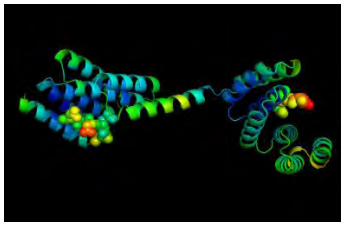
1.5.4 In-depth Review of the Structure of Hop

A peer-reviewed overall structure for Hop has not been published to the Protein Data Bank (PDB) for any species as yet, however there have been multiple studies done to discern Hsp70:Hop and Hsp90:Hop complex structure (Lee, Graf, Mayer, Richter, & Mayer, 2012a; Romano et al., 2009; Schmid et al., 2012; Southworth & Agard, 2011a). These studies have isolated and determined several domains, whose individual structures have been solved via X-ray crystallography,nuclear magnetic resonance (NMR) and cryogenic electron microscopy (Cryo-EM) experiments.

1.5.5 Three TPR Domains

There are three, well-conserved and very similar functional Tetricopeptide Repeat (TPR) domains (Lee et al., 2012); TPR1 and the tightly linked TPR2A and TPR2B in both yeast and human Hop (see Table 1.2). The N-terminal TPR region (TPR1) bindsthe EEVD C-terminal Hsp70 peptide motif (Chen et al., 1996; Lässle et al.,1997; Chen and Smith, 1998; van Der Spuy et al., 2000) and Hsp104 (Abbas-Terki et al., 2001); the TPR2B domain also binds the EEVD C-terminal residues of Hsp70, as well as the Hsp90 M domain (Scheufler et al., 2000; Southworth & Agard, 2011a) and the TPR2A domain binds the MEEVD C-terminal residues of Hsp90 (see Figures1.2 and 1.3).

Table 1.1: TPR regions involved in the functioning of Hop.

Complex and Organism	Image	Contact Residues	PDB ID and Reference
Hsp90 Mdomain and TPR2A&B Yeast		In the complex, TPR2A is oriented towards the C-terminal part of Hsp90-M directly contacting the outside of Hsp90-M only with residues of helix 7 (residues 368–374).	Unpublished structure produced by homology modelling (Schmid et al., 2012)
Hsc70 c-terminal and TPR2A Yeast		Lys- 301 and Asn-298 Lys-229, Asn-233, and Asn-264, Arg-305	3ESK (Kajander et al., 2009)
Hsp90 c-terminal and TPR2A Human		Lys 229, Asn 233, Asn 264, Lys 301, and Arg 305	1ELR (Scheufler et al., 2000)
Hsc70 c-terminal and TPR1 Human		Lys 8, Asn 12, Asn 43, Lys 73, and Arg 77	1ELW (Scheufler et al., 2000)
Hsp70 c-terminal and TPR2B Yeast		Lys 229, Asn 233, Asn 264, Lys 301, Arg 305,	3UPV (Schmid et al., 2012)
Hsp90 C-terminal and Hsp70 C-terminal and TPR2A&B Yeast		See Figure 1.3.	3UQ3 (Schmid et al., 2012)

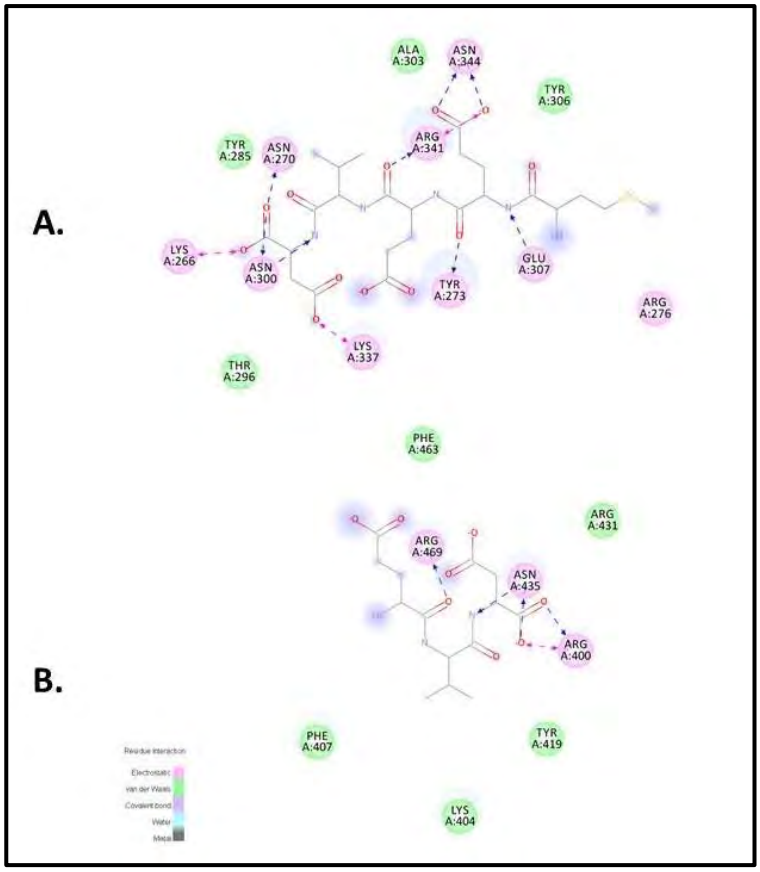


Figure 1.3: Concave surface interaction in ScHopTPR2 A) TPR2A, showing active residues forming the binding site with C-terminal MEEVD motif in Hsp90. B) TPR2B, showing active residues forming the binding site with C-terminal EVD motif in Hsp90 (images produced in Discovery Studio Visualizer–Accelrys Inc., 2009).

This EEVD motif has a general pI of 3.4247. However there is an isoform of PfHsp70 (PfHsp70-x (Hatherley, 2012)) that significantly differs itself from human, yeast and three other PfHsp70 isoforms by instead possessing, among other anomalous features, a C-terminal EEVN Motif, which possesses a general pI of 3.6136 (calculated in MATLAB, using the bioinformatics toolkit). This may indicate a functionally significant difference between HsHsp70:HsHop binding and PfHsp70-x:PfHop binding.

1.5.6 Two DP Domains

The Hop protein possesses two structurally similar domains; DP1, which lies between TPR1 and TPR2A; and DP2, positioned between TPR2B and the C-terminal end of Hop (Schmid et al., 2012). The alignment for these domains in yeast is displayed in Figure 1.4.

```
Identities = 18/75 (24%), Positives = 44/75 (59%)
01  QPDLGLTQLFADPNLIENLKKNPKTSEMMKDPQLVAKLIGYKQNPQAIGQDLFTDPRLMTIMAT
    ||  | ::   : :  : : |:|::: :|:|| : : |   :||| |: |: : :|:::  : |
01  QP--GTSNETPEETYQRAM-KDPEVAAIMQDPVMQSILQQAQQNPAAL-QEHMKNPEVFKKIQT

65  LM--G-VDLN-
    |:  | :  :
61  LIAAGIIRTGR
```

Figure 1.4: Global alignment of the sequence data for PDB entries 2llw and 2llv.

The DP regions are so-called owingto the well conserved DP (Aspartic acid and Proline) repeats within these structures (Odunuga et al., 2004; Schmid et al., 2012). Recent structural studies in yeast show that DP domains and especially the DP2 domains are important for Hsp90 client processing in yeast (Schmid et al., 2012). The NMR crystal structures associated with this study are shown below (viewed in PyMOL, Figure 1.5):

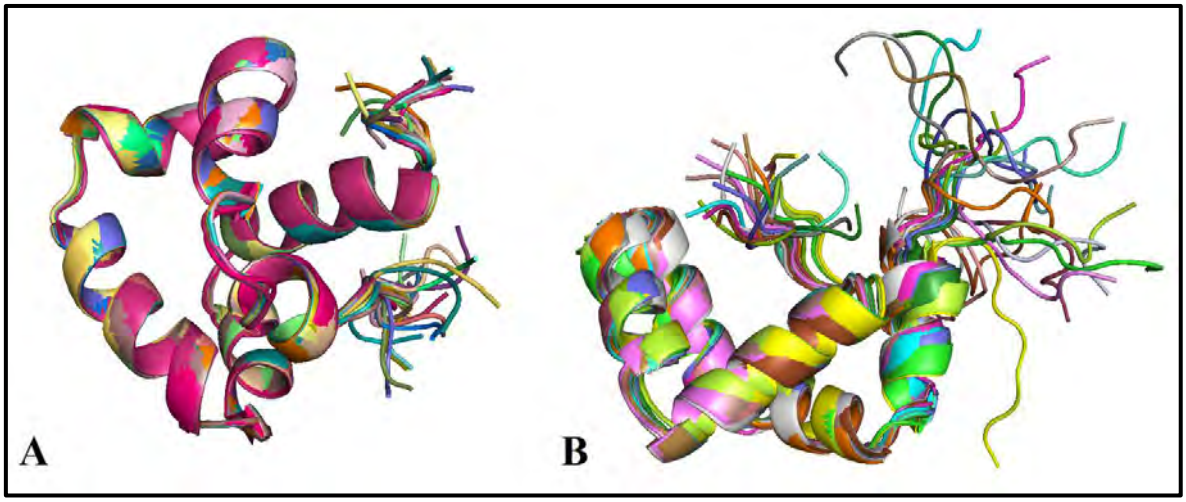


Figure 1.5: The superposed, 21 NMR solution models for ScHopDP domains. A) DP1 (PDB entry: 2LLV) and B) DP2 (PDB entry: 2LLW). The models appear to be in good agreement for DP1. DP2 models were also in good agreement, but more flexible.

From Figure 1.5, it is not plain to see, but DP2 is comprised of five helices, forming an elongated v-shape. DP1 is comprised of roughly the same five helices, but with a short additional helix near the n-terminal, and is more globular than DP2. This difference has been noted in the literature, and DP1 is reported as being ‘denser’ than DP2 (Schmid et al., 2012).

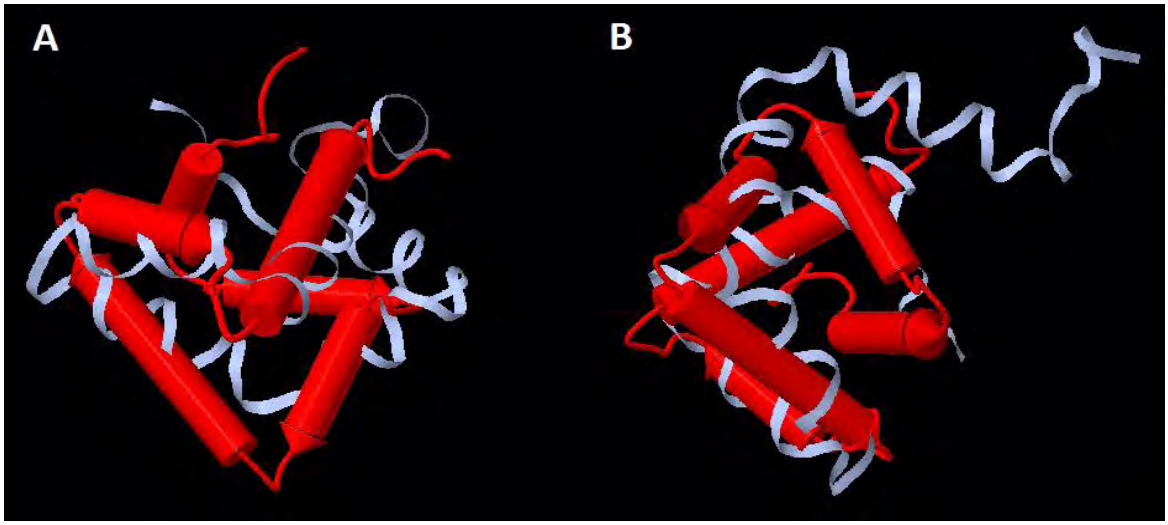


Figure 1.6: DP1 (red rockets) and DP2 (grey ribbons) superposed. A) The four C-terminal helices (foreground) coincide relatively closely for the two structures. B) A 180° rotation about the horizontal axis in A. This view shows where the alignment of the two DP regions ends; the N-terminal segments seem to diverge from the fifth helix, which stretches away from the main body of the domain in DP2, but for DP1, folds in on itself, almost burying the additional sixth helix inside the protein.

To examine the differences in structure both structures were aligned, superimposed and visualised in molviewer (The MathWorks Inc., 2009), see Figure 1.6. As can be seen from both Figure 1.4 and 1.6, the four C-terminal helices align better and relatively closer for the two structures than do the N-terminal segments, which diverge from the fifth helix onward. These stretch away from the main body of the domain in DP2, but for DP1, folds in on itself.

Native	D	P	E	V	Q	Q	I	M	S	D	P	A	M
APAV	A	-	A	-	-	-	-	-	-	-	-	-	-
APAM	-	-	-	-	-	-	-	-	-	A	-	-	-
AP2	A	-	A	-	-	-	-	-	-	A	-	-	-

Figure 1.7: Three DP2 Alanine substitution mutant types studied by Chen and Smith (2008).

The exact functionality of the DP regions is not yet entirely clear, and there are conflicting results with respect to the effects of mutations in DP2. Chen and Smith (2008) found that Alanine substitution in the relatively well conserved DPEV motif (Odunuga et al., 2003) region of DP2 (see Figure 1.7) disrupts interaction with Hsp70 possibly by perturbing some inter-domain interaction or structural integrity in certain strains of yeast. However,

Flom et al (2007) found that complete deletion of DP2 (let alone point mutation within DP2) did not inhibit the Hsp70 interaction in other strains. It is probable that these conflicting results reflect yeast strain-specific differences.

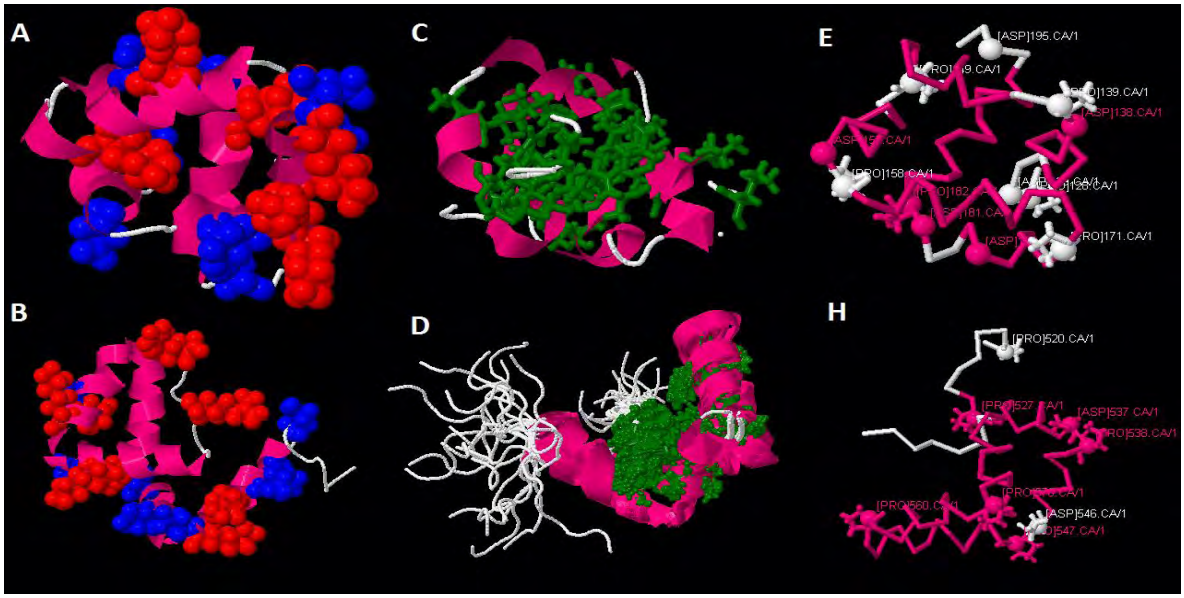


Figure 1.8: The amino acid property views for DP2 (bottom, PDB entry: 2llw) and DP1 (top, PDB entry: 2llv). A&B) Amino Acid Charge, Positive = red, Negative = blue. C&D) Hydrophobic Amino Acids = green. E&F) Aspartic Acid and Proline alpha carbons labelled and highlighted.

Other properties of these two proteins were compared in molviewer (see Figure 1.8). In general both possess an equal distribution of both negatively and positively charged amino acids, and hydrophobic residues are positioned internally. Both possess DP-repeats between the helix regions.

1.5.7 Nuclear Localisation Signal (NLS) Regions

A short bipartite Nuclear Localisation Signal (NLS) exists, overlapping the C-terminal edge of TPR2A (Daniel et al., 2008; Longshaw et al., 2004). A second, putative NLS may exist in the DP2 domain, although its functionality has yet to be shown (Odunuga et al., 2004).

1.6 Current Antimalarial Hsp Drugs

Hsps are currently being studied as potential drug targets for a range of diseases such as cancer (Odunuga et al, 2004) and neurodegenerative diseases (Romano et al., 2009). Table 1.1 highlights a list of Hop and Hsp90 inhibitors.

Table 1.2: A selection of currently researched potential inhibitors of Hsps interacting with Hop

Inhibitor	Structure/complex	Binding/inhibition site	Reference
Celastrol	Cdc37:Hsp90	Hsp90 C-terminal	(Zhang et al., 2009)
Cytotoxic sugars	Hsp90	Hsp90 C-terminal	(Donnelly et al., 2010)
Biotin-related	Hsp90:HopTPR2a	Hsp90 C-terminal	(Yi et al., 2009)
"AntP-TPR" peptide	Hsp90:HopTPR2a	Competes with Hop for Hsp90 C-terminal	(Horibe et al., 2011)
Novobiocin and Couermycin	Hsp90	Hsp90 C-terminal	(Matts et al., 2011)
Gealdinimycin and 17AAG	Hsp90	Hsp 90 ATP Binding Pocket	(Kumar, Musiyenko, & Barik, 2003)
Sansalvamide A and analogs	Hsp90	N-terminal and Middle Domain	(Kunicki et al., 2011)
Prion Protein Fragments	Hop:prion	Hop TPR2a	(Romano et al., 2009)

1.7 Possibility for Human and Malarial Hsp Interaction

While the mature erythrocyte does not possess a nucleus or mitochondria, and therefore does not synthesise new proteins, analysis of the red blood cell proteome does however show that the mammalian (at least, in mice and human) red blood cell maintains approximately 700 proteins throughout its 120 day lifetime (Pasini et al., 2008) including two cytosolic Hsp70s and a cytosolic Hsp90 (Gromov & Celis, 1991). While there is no direct experimental evidence for the presence of Hop in the mature red blood cell (and this is probably because no one has gone looking for it), the presence of Hsp70 and Hsp90 (especially considering the role Hop plays coordinating these Hsps) and other Hsps (Pasini et al., 2008), suggest that human Hop could be present in the red blood cell.

If the above conclusion is correct, and in light of a recent finding which show that a PfHsp70 variant (PfHsp70-x) is transported to the host erythrocyte (Kulzer et al, 2012); it is possible that cross-species interaction between PfHsp70-x (and possibly even PfHsp90) and HsHop could occur. As yet, this possibility has not been investigated experimentally or *in silico.*

1.8 Overall Research Rationale for the Project

The results of the Human Genome Project (and other genome projects since, e.g. Brayton et al. (2007) and Cornillot et al. (2012)) have provided many potential drug targets that were once hard to come by, and the pharmacological industry is now left with the challenge of mining genomic data in search of the proteins that will be most effective in fighting human disease (Smith, 2003). Although new data is being published every day, relatively few of the known 35,000 genes in the human genome have described functions (Overington et al., 2006). The constant rate at which drugs against new proteins are launched starkly contrasts the significantly lower rate of developing drugs against new protein families (Overington, Al-Lazikani, & Hopkins, 2006).

Two Hsps, Hsp90 and Hsp70, are currently the subjects of intense research to find new drugs and their corresponding drug-targets in the fight against increasingly drug resistant *P. falciparum*. Unfortunately, Hsps are well conserved in most species, thus human and *P. falciparum* proteins are very similar. Careful computational analysis of the proteins will save time when it comes to determining the feasibility of testing new drugs and their protein targets in the wet lab. A lessor understood co-chaperone, the Hsp70/Hsp90

organising protein, has been found to play an essential role in modulating the activity and co-interaction of these two essential chaperones. The best understood aspects of Hop so far indicate that three Tetratricopeptide repeat (TPR) domains in the protein bind to specific C-terminal motifs in Hsp70 and Hsp90.

Developing a drug against a particular target is an expensive and high-risk investment (close to a billion US dollars). Once a drug target enters a pharmaceutical company's research and development phase, it may take up to 12 years to develop the final marketable product, assuming it makes it to that point (Smith, 2003). Clinical trials disappoint for two basic reasons: drugs don't work or they turn out to be unsafe. This is primarily as a result of errors in the validation process (Smith, 2003). Most drugs are inhibitors that block the action of a particular target protein, and may cause disruption in the human cell if they inhibit other non-target proteins essential to cell functioning (Ma & Nussinov, 2007). The only way to be completely certain that a protein is instrumental in a given disease is to test the inhibition of that protein in humans, however, its role in disease must be clearly understood before it is used to screen for drug susceptibility, let alone before human trials (Overington et al., 2006).

So far, protein structures for both the Hop TPR domains and the respective C-terminal motifs for Hsp70 and Hsp90 in complex have been published for two species (*Homo sapiens* and *Saccharomyces cerevisiae*). Recent research suggests that Hop, Hsp90 and Hsp70 do in fact interact and form complex in the *P. falciparum* trophozoite (within the infected host erythrocyte) and is overexpressed in this infective stage. While it has been found that *P. falciparum* transports parasite proteins to the host erythorocyte cell, it has only been suggested that it may be doing the same for its Hsp's. However, recent data show that a PfHsp70 variant (PfHsp70-x) is transported to the host eythrocyte. It has been suggested that PfHsp's may interact with human Hsps, effectively annexing the host cell proteins for parasite protein refolding. In this case, there is potential for complex to form between PfHsp70-x and human Hop. However, there has been almost no computational research on malarial Hop protein in complex with other malarial Hsps, let alone human Hop with malarial Hsps.

1.9 Aims

The current work was undertaken to focus on several aspects of the *in-silico* characterisation of PfHop. The aims of this work were to yield a valuable starting point to understand the scope of both the variability and conservation of several domains, motifs and residues on which functional interaction studies have been centred in PfHop and HsHop, in order to identify regions of the protein that could potentially function as drug target sites. In order to do this, the primary objectives of the work were to build comparable homology models of Hop in complex with several of its protein/peptide partners in both human and *P. falciparum* and compare the interactions involved in complex formation.

Chapter 2: Multiple Sequence Alignment and Phylogenetic Analyses

2.1 Introduction

Earlier this year, *P. falciparum* Hop was localised and isolated from the trophozooite (infective) stage of the parasite (Gitau et al., 2012). This analysis included a very basic multiple sequence alignment of the protein (which included an error in sequence and taxon identity), leaving much to be determined with regards to understanding this protein's phylogeny. An earlier review by Odunuga et al (2008) included a multiple sequence alignment that was very thorough and from a wider range of taxa. However, considering recent advances in determining the structure of Hop, a new in-depth multiple sequence alignment of Hop protein with representatives from four Kingdoms (Animalia, Fungae, Plantae and Protozoa), fifteen major taxonomic groups and 60 species was undertaken as part of the current work.

2.2 Multiple Sequence Alignment

The aims of this project were to compare two species' Hop proteins and determine how different they are from each other, in order to determine the feasibility of targeting this protein for drug research. However, it is only possible to completely gauge how different two sequences are within the context of a scale of variability (Pei et al., 2008). This is the premise for multiple sequence alignment (MSA).

In most cases, alignments should be done at the protein level. One exception is when there has been a frame shift mutation in one of the sequences such that the amino acid sequences differ significantly but the DNA sequences coding for those proteins can still be aligned. The optimal choice also depends on the level of evolutionary relationship being investigated, as well as the purpose of the research. If closely related species/strains are being analysed, then DNA analysis will be more informative as it allows detection of synonymous changes (Harrison & Langdale, 2006). If more distant evolutionary relationships are being studied, then analysis of protein sequences is more appropriate because the protein sequences change more slowly.

2.2.1 BLAST

The Basic Local Alignment Search Tool (BLAST) was originally designed to detect homologous nucleotide sequences of a query sequence by directly approximating alignments that optimize a measure of local similarity; the Maximal Segment Pair (MSP) score (Altschul et al., 1990). It has since been implemented and optimised within an online public interface (Johnson et al., 2008) to run on large-scale globally accessible protein and nucleotide sequence databases such as NCBI and GENBANK and their respective curated versions (Pruitt & Maglott, 2001). Several algorithms have been further developed for homolog-searches using different types of sequences with greater efficiency (Altschul et al., 1997; Price, Dehal, & Arkin, 2008). It is currently one of the most widely used tools for conducting bioinformatics research and is an excellent resource for teaching some of the foundational principles of bioinformatics (Kerfeld & Scott, 2011).

Protein BLAST requires the submission of a query sequence to use for searching selected databases. Once matches have been found they are aligned and scored according to the Expect value (E), which is a value that estimates how often one will find false matches by chance when searching a database of a particular size. The closer the E-value is to zero, the more significant the match is and it decreases exponentially as the MSP of the match increases. The E value takes account for the length of the query sequence, because shorter sequences have a higher probability of occurring in the database by chance.

A good place to start when looking for conserved domains is the NCBI domain recognition tool when performing a pBLAST against the query of interest. When *P. falciparum* was pBLAST against the non-redundant RefSeq database several domains were recognised (Figure 2.1).

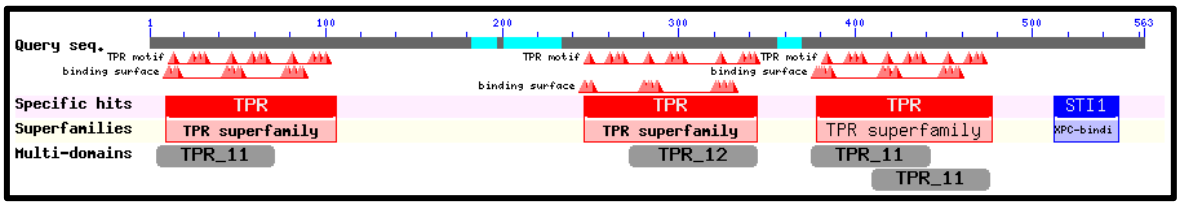


Figure 2.1: Domains in the Hop protein sequence recognised by NCBI pBLAST tool.

The three TPR repeat motif domains are recognised as they are a common feature in several proteins and have been extremely well characterised by the scientific community. The two DP domains were not recognised most likely because their structures were only

elucidated in 2012 and a lot of work remains to be done with regards to the characterisation of these domains. Interestingly, an XPC (xeroderma pigmentosum group C protein) binding domain was recognised in the region that was expected to contain the DP2 domain. This domain is part of a protein called hHR23, belongs to a protein superfamily designated cl15287. Members of this family adopt a structure that is very similar to that describing DP2 (see Figure 2.2), consisting of four alpha helices, arranged in an array. They bind specifically and directly to XPC to initiate nucleotide excision repair (NER) (Kamionka & Feigon, 2004). The protein is of interest owing to its role in a very rare autosomal recessive genetic disorder known as xeroderma pigmentosum (XP), whereby afflicted individuals are NER deficient. These individuals often die very young owing to cancers developed through disruption of the oncogenes (that are not repaired) during exposure to UV light (Kamionka & Feigon, 2004).

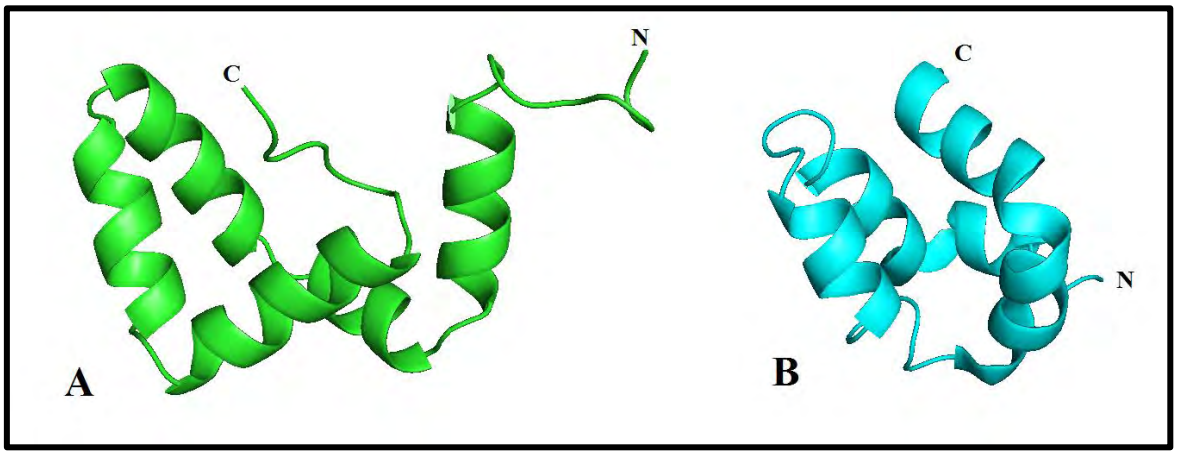


Figure 2.2: BLAST identified DP domain structures. A) ScHop DP2 (PDB ID: 2LLW). B) Human hHR23 XPC binding domain (PDB ID: 1OQY).

The XPC binding domain (Figure 2.2B) bears some structural similarity to the DP2 domain but appears more globular, like ScHop DP1. Hydrophobic interactions between the C-terminal helix and the rest of the protein are most likely the reason for this globular structure, while hydrophilic patches on the outer surface of the protein are reportedly responsible for domain functioning (Kamionka & Feigon, 2004). This is very similar to both ScHop DP domains, where hydrophobic residues are interior to the protein (see Figure 1.8D and B).

2.2.2 COBALT

COBALT performs local alignment in a constraint-based way by looking for conserved domains and forming pairwise alignments around these regions and then constraining the rest of the multiple sequence alignments in a progressive sequencing manner. It works slowly, but the algorithm appears to function more accurately if there are conserved regions in the sequences being aligned (Papadopoulis and Agarwala, 2007).

2.2.3 MAFFT

MAFFT is a MSA program that uses fast Fourier transforms (FFT) to convert the amino acid sequence in question to a sequence of vectors, representing the volume and polarity of each residue. This is because evolution favours amino acid substitutions which retain similar physico-chemical properties (Jones, Taylor, & Thornton, 1992) and the MAFFT program is then capable of scoring residue pairs according to their vectors. 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 with greater efficiency than other widely used MSA programs. Owing to its wide use and popularity, MAFFT has been developed to utilise several different alignment strategies optimised to specific types of data (Katoh & Toh, 2008). MAFFT's E-INSi protocol has been developed to tackle proteins such as RNA polymerase, which has several conserved motifs embedded in long, unaligned regions. This is a useful feature that allows distantly related proteins with one or more long, non-conserved regions to be aligned. This algorithm must be used with caution as it assumes that the arrangement of the conserved domains is shared by all sequences (Katoh *et al*., 2005). For proteins with global homology, the G-INSi protocol uses iterative refinement to improve the accuracy of the alignment. This is done using the WSA scoring method, based on a matrix created by analysing only the well conserved segments of sequences (Katoh *et al*., 2005).

2.3 Motif Analysis

Multiple Expectation Maximisation for Motif Elicitation (MEME) is a tool for identifying biologically functional motifs in a group of related DNA or amino acid sequences (Bailey and Charles, 1994). It does this by discovering motifs in a collection of sequences through using expectation maximisation (which enables parameter estimation in probabilistic models with incomplete data (Do & Batzoglou, 2008)) to fit a two-component finite

mixture model (Figueireido, 2002) to the sequence set. The tool can estimate how many times each motif (even with differing numbers of occurrence per sequence) occurs on each sequence in an unaligned dataset. It returns this motif as a sequence logo with several scores (Bailey and Charles, 1994; see also Appendix 5). The E value for each motif is an estimate of the expected number of motifs with the log likelihood ratio of the returned motif (with the same width and site count) that one would find in a set of random sequences of the same number as the input set (Bailey and Charles, 1994).

2.3.1 MEME Block Diagrams:

Each sequence is displayed as a line overlain with block diagrams in colours that represent specific found motif numbers, and scored according to the positional *p*-value (Bailey and Gribskov, 1998). The *p*-value of a sequence is computed from the score generated by matching the motif site/s on any given sequence with the position specific scoring matrix for the motif. The *p*-value returns the probability of a random sequence having an equivalent match score or higher (Bailey et al., 2009). The height of the block representing a motif site gives an indication of the significance of the match; the height is proportional to the negative logarithm of the sequence's *p*-value, cut off at the height for a *p*-value of $1e^{-10}$. As such, taller blocks represent motifs that are more significant (Bailey et al., 2009).

2.3.2 MEME Sequence Logos

MEME motifs are represented by position-specific probability matrices that dictate the probability of each possible letter appearing at a specific position in all occurrences of the motif (Bailey and Charles, 1994). This logo contains a stack of letters at every position in the motif. The height of the individual letters in a stack is the probability (in bits) of the letter at that position multiplied by the "total information content" (the number of times that residue occurs within that residue site in each motif site in the dataset) of the stack (Bailey et al., 2009). Thus the total height of the stack is reduced if the residue site is not well conserved at that position in the motif.

Table 2.1: Amino acid colour codes for sequence logos in MEME.

Amino acids	Colour	Properties
A C F I L V W M	Blue	Hydrophobic
N Q S T	Green	Polar, non-charged, non-aliphatic
D E	Magenta	Acidic
K R	Red	Positively charged
H	Pink	Positively charged, cyclic sidechain
G	Orange	Simple, non-polar
P	Yellow	Cyclised
Y	Turquoise	Non-polar, aromatic

For proteins, the colours of the individual letters in the motif are based on the biochemical properties of the various amino acids according to the convention in Table 2.1 (from Kyte and Doolittle, 1982).

2.3.3 MAST

MAST jobs can be completed in parallel to MEME jobs within the MEME Suite Webserver (Bailey et al., 2009). MAST determines the probability that two motifs are significantly different by calculating the pairwise correlations between each pair of motifs. The maximum, found by trying all alignments of the two motifs, is the sum of Pearson's correlation coefficients for aligned columns divided by the width of the shortest motif in the pair (Bailey and Gribskov, 1998). Pairs of motifs with higher correlations are too similar and should not be considered as separate motifs. The server returns a pairwise "similarity" table where correlations above the similarity threshold (i.e. with lower significant difference) are shown in red text. Additionally, the server returns a top scores table, where each of the sequences in the original dataset are displayed as lines, overlain with the original block representation of motifs (as for MEME block diagrams) and ranked according to the lowest MAST E-value. The MAST E-value is equal to the combined position *p*-value of the sequence times the number of sequences in the database (Bailey and Gribskov, 1998). More simply, it is the expected number of sequences in a random database of the same size that would match the all motifs on the sequence in question as well as the sequence itself does.

2.4 Phylogenetic Analysis

MSA is also an incredibly important step in protein structure analysis; it allows the researcher to identify well conserved domains (and thus most likely functional) and important residues (such as those involved in binding ligands or other proteins). It has become well established practise to use MSA and phylogenetic analyses to help predict the structure of unknown proteins, as conserved functional domains tend to have conserved folds, particularly within those domains (Benner, 2001). Phylogenetic analyses have been found to effectively assist in predicting unknown gene function, which is the backbone of current drug discovery, by tracing genes of known function and comparing how they are related to unknown genes (Searls, 2003).

2.4.1 Species Trees versus Protein Trees

Species trees (Phylogenetic trees) form a pattern of branching of species lineages via the process of predicted or inferred speciation. Gene/protein Trees are formed through the inferred mutation and recombination events and seem to be broken into several pieces

owing to recombination within populations. Protein trees are usually contained within the branches of species phylogeny. If in agreement, protein sequences should show the same branching topology as a species tree, but may have more terminals when analogous and orthologous genes encoded for the proteins are considered (Harrison & Langdale, 2006). This is because genes can undergo recombination, insertion, deletion within species and do this several times. Genes can also break the confines of species lineages in other ways. A good outgroup may allow the tree to indicate true evolutionary direction.

In the absence of a suitable outgroup, the root may be positioned by assuming approximately equal evolutionary rates over all the branches. In this way the root is put at the midpoint of the longest pathway between two operational taxonomic units (Harrison & Langdale, 2006). This is the molecular clock method, which has other problems associated with it (as it is based on the assumption that all sequences have equal evolutionary rates).

2.4.2 MEGA

MEGA is a software package that equips biologists with all major phylogenetic analysis tools (Kumar, Tamura, & Nei, 1994). With the latest release, the collection of analysis tools in MEGA now includes the maximum likelihood (ML) methods for molecular evolutionary analysis (Tamura et al., 2011). Phylogenetic inference from amino acid sequence data uses mainly empirical models of amino acid replacement and has therefore been dependent on the Dayhoff and JTT amino acid substitution models (Jones et al., 1992), and other more recent models (Whelan and Goldman, 2001). Maximum likelihood analyses calculate the probability that a data set fits a tree derived from that data set, given a specified model of sequence evolution (analogous to models of amino acid substitution used to find homologous sequences). To select a model of sequence evolution, the data must be roughly compared against a set of models of sequence evolution, then the model that best describes the observed pattern of sequence variation (has a higher optimality criterion) is selected for further analyses (Harrison & Langdale, 2006).

2.4.3 Evolutionary Model Selection

Model selection is well established for making inferences from observational data, especially when data are collected from complex systems (Harrison & Langdale, 2006). This is critically important when addressing evolutionary research questions where experimental manipulation is not possible. In general, the researcher selects a model by scoring the model to the data with some form of optimality criterion. The optimality criterion for the correct model describing the structure of a phylogenetic tree is a score calculation based on the substitution of characters (such as nucleotides or amino acids) to explain the observed sequences at the terminal branches of a tree (Harrison & Langdale, 2006). To calculate such a score, one needs some measurement of change. Observed difference is often used but may underestimate the degree of actual change. For example, if there are three increments of change, one character may have changed once and then back to its original state again (back mutation). In this case, there will be no observable difference, but there have been two changes and such back mutations need to estimated.This is usually done with a network such as a hidden Markov model. The optimal model will depend on the kind of data that is being used to construct a tree, i.e. what the characters and character states are. If one uses protein sequences, the characters may be sites and the states the different amino acids.

Model selection programs in phylogenetic application packages are commonly used to select appropriate models based on aligned sequence data. MEGA5 provides the goodness-of-fit test of the substitution models with and without assuming the existence of evolutionary rate variation among sites. The goodness-of-fit of each model to the data is measured by the Bayesian information criterion (BIC, Beaumont & Rannala, 2004) and a corrected revision of the Akaike information criterion score (AIC^c).

Both of these criterion indicators persistently select substitution models that are more complex than the true model. However, the true model usually appears among the top 3 when BIC was used and among the top 5 when AIC^c was used (Tamura et al., 2011). As such, several maximum likelihood trees should be built with a variety of models, and all resulting trees compared to their bootstrapped trees as well as to each other, in order to find the most correct tree.

2.4.4 EnsEMBL Compara-Gene Trees

EnsEMBL (Kersey et al., 2010) trees for a specific gene of interest are based on maximum likelihood phylogenetic trees built by TreeBest (Li, 2006) utilising the TreeFam Database. TreeFam (Li, 2006; Li et al., 2006) was developed to provide curated phylogenetic trees for all animal gene families, as well as automatically predict orthologs and paralogs. These trees, representing the evolutionary history between the genes, are then compared to their species trees in order to differentiate between duplication and speciation events (Li & Durbin, 2007). The latest release of TreeFam contains curated trees for 1314 families and automatically generated trees for another 14351 families from 25 fully sequenced animal genomes, as well as four genomes from plant and fungal outgroup species. EnsEMBL utilises TreeFam and TreeBest to automatically group genes into families and for building phylogenetic trees (Kersey et al., 2010) and thus these trees are not without error and must be used with caution. The gene tree for a particular gene is accessible via the left-hand navigation menu of the gene's page on EnsEMBL, along with ortholog and paralog data which are available for querying via the BioMart interface and/or a Perl API (Kersey et al., 2010; Ruan et al., 2008).

2.5 Methods and Software

2.5.1 Sequence Retrieval

Sequences for approximately 70 organisms from a wide range of taxa were acquired from the NCBI (Tatusova et al., 2012) and PlasmoDB.org Databases. *P. falciparum* Hop protein was downloaded from PlasmoDB and then used for protein BLAST against every eukaryote genome available in the NCBI BLAST's Genome Map viewer. All sequences with an E value > e^{-50} were downloaded and saved to *.fasta* format files. A summary of all sequences originally attained is presented in Appendix 1. The high E value was selected to filter out non Hop-homologs with the TPR regions, as this motif is a common (and highly conserved in structure and binding site composition) motif in other proteins such as FKBP51 and FKBP52 (Cheung-Flynn et al., 2003). Sequence representatives were found from four Kingdoms and several Phyla. No prokaryote sequences were analysed as the primary focus of this study was eukaryote organisms (i.e. *H. sapiens* and *P. falciparum*). Interestingly, there were no suitable homologs of Hop in the three avian species searched (Turkey, Chicken and Zebra Finch) as well as the single Marsupial (*Monodelphis*

domestica) species searched. To account for the possibility that protozoan and avian Hop proteins are too diverse, another vertebrate species' Hop sequence (*H. sapiens*) was used to BLAST the genomes of the three avian species, still returning no Hop homologues with an Evalue > e^{-50}. An NCBI search for "Hop/Sti1" protein for "Avian/Aves/*Gallus gallus*" returned a poor homologue under the title of "hsc70-interacting protein St13". To check for the likelihood that all sequences shared homology to all other sequences, a short script (see Appendix 2, Section A) was written to filter through a .fasta file of pairwise of alignments of all sequences (produced in Jalview) for a minimum sequence identity of 30% (Rost, 1999). This resulted in selection of the final 60 sequences to be used for MSA.

2.5.2 Multiple Sequence Alignment

All 60 sequences were initially aligned in COBALT, a relatively new sequence alignment algorithm that has recently been promoted on the BLAST webpage. A second alignment was produced using MAFFT's E-INSi's protocol. Hop possesses two problematic regions making the selection of this protocol optimal: the linker region between DP1 and TPR2A and that between TPR2B and DP2. From looking at all MSA results, it is possible to see that these regions possess the most gaps and unaligned regions. A second MAFFT protocol, the L-INSi method produced slightly less comparable results. These results were compared with a further alignment program DIALIGN-TX (Subramanian, Kaufmann, & Morgenstern, 2008), using two different parameter sets: default and increased width of low conserved regions expected to be found. A final MSA method for the whole protein was settled on (utilising the E-INSi protocol) and this was then used for phylogenetic analysis.

2.5.3 Meme Whole Protein Analysis

Sequences were sent to the MEME webserver (Bailey et al., 2009) to search for conserved motifs/domains in sequences. The unaligned sequences were submitted to the MEME server, with the following parameters: motif width to be searched for was between 2-150 residues; motif can occur any number of times in a sequence (this is owing to previous research finding TPR1 and TPR2B to be very similar in structure) and site limitations were left empty (this is the default, i.e. no sequences are excluded from the analysis). This approach was used each time in subsequent MEME analysis. These parameters were kept constant when searching for 5, 10, 15 and 20 motif sets in all 60 sequences.

2.5.4 Domain Analysis

Once a final MSA method for the whole protein was settled on (MAFFT E-INSi protocol), whole Hop was realigned (MAFFT E-INSi) with fragments of Hop protein relating to well-studied structural domains in human and yeast, namely the TPR1, DP1, TPR2A&B and DP2 domains. These fragments were obtained by retrieving the protein sequences from several published structures of these domains on the PDB; namely 1ELW (TPR1), 2LLV (DP1) 3UQV (TPR2A&B) and 2LLW (DP2). This strategy highlighted regions of good alignment for most domains (except DP1). Once realigned, the fragments were removed and the four regions roughly corresponding to each domain were cut out of the alignment in Jalview (Waterhouse et al., 2009) and saved separately. The three TPR motif and DP2 motif regions were resent to MAFFT for global realignment using the G-INSi protocol. DP1 was realigned using the L-INSi protocol and E-INSi protocols. The final aligned domains were also used for phylogenetic analysis.

2.5.5 MEME Domain Analysis

These domain datasets were then also submitted to MEME using the parameter set listed above, however, making the max sequence length for that dataset the max length of the motifs to be searched for. Only 15 motifs were sought per domain.

2.5.6 Phylogenetic Analysis

The protein phylogenies were created in MEGA. AIC^c and BIC scores were used to determine the correct evolutionary models to use for maximum likelihood analysis. The overall phylogeny was selected from the top three models selected from lowest BIC scores. Both slow, Maximum Likelihood (ML) and fast, Neighbour-Joining (NJ) tree-building algorithms were used for the full-length Hop Protein analysis. Three different evolutionary models were tested in the construction the full-length Hop ML protein tree, rtREV, JTT and WAG, however owing to the excessively long run time to build ML trees using the WAG and JTT models, these methods were only used for comparison of the full-length protein.

2.6 Results

2.6.1 Sequence Selection

See Appendix 1 for all pBLAST results. Exactly 60 out of approximately 70 sequences were selected for final MSA analysis. These included representatives from the vertebrates, apart from the aforementioned avian and marsupial species. Additionally the *A. gambiae* species sequence was not utilised do to a lack of information in the database as to the origins of the sequence. Using the script in Appendix 2, Section A it was determined that these final 60 sequences all shared 30% sequence identity with every other sequence.

2.6.2 Multiple Sequence Alignment

The full-length protein MAFFT E-INSi alignment was viewed with Jalview. Several coloration options were used to get an overview on certain properties of the protein that are of significance to the current work. In Figure 2.3, it is easily seen from the first three panels that the regions relating to DP1 and the linker regions possess the least conservation. These regions are also the least well aligned regions, and the sites of most insertions and deletions. The exact residues relating to these regions will be analysed in depth in Section 2.6.4.

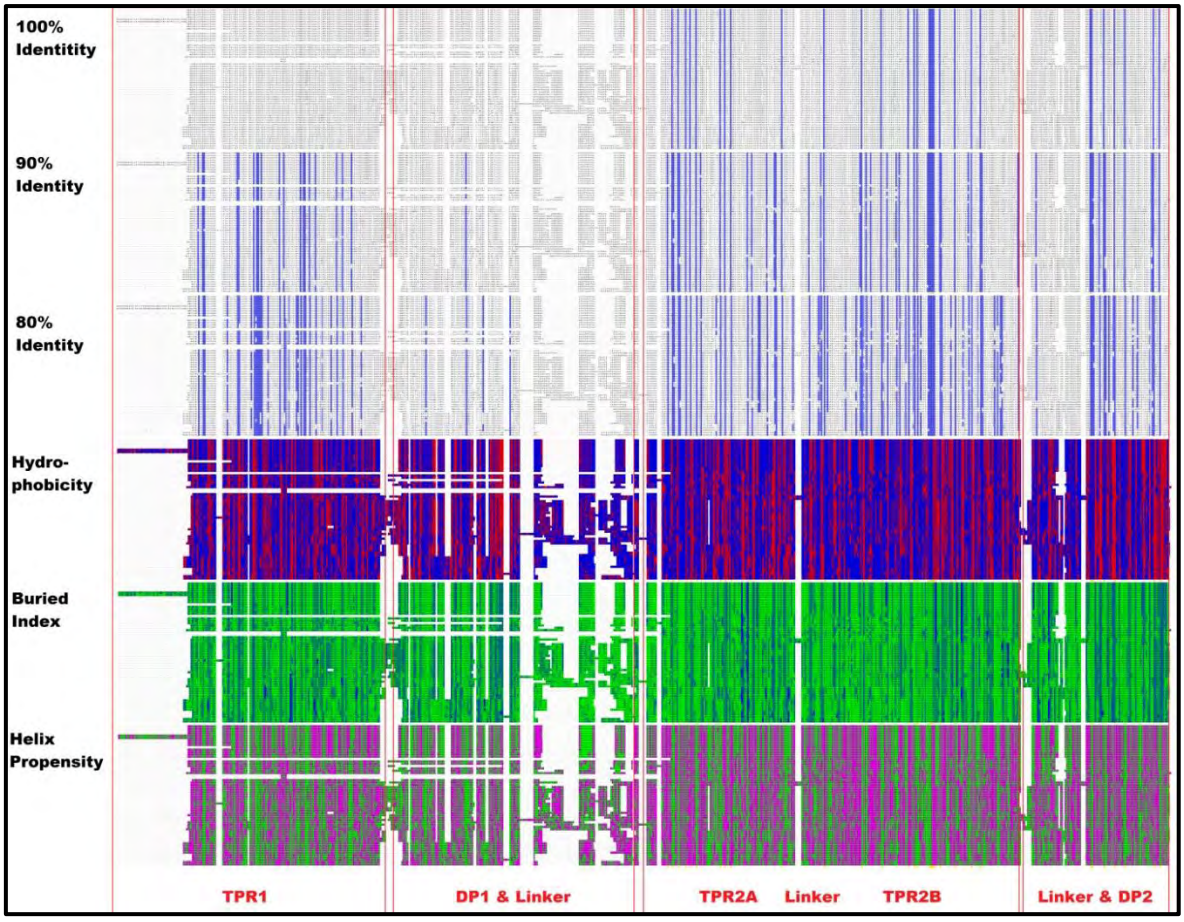


Figure 2.3: General features of Hop Protein for all sequences analysed. The residues in the first three panels are coloured by Blosum62 scoring if they are above the respective identity threshold. *Blosum62*: Gaps are coloured white. If a residue matches the consensus sequence residue at that position it is coloured dark blue. If it does not match the consensus residue but the two residues have a positive Blosum62 score, it is coloured light blue. *Hydrophobicity*: Amino acids are coloured according to the hydrophobicity table of Kyte and Doolittle (1982). The most hydrophobic residues according to this table are on the red end of the spectrum and the most hydrophilic ones are on the blue end. *Buried index*: according to Jalview, residues in the dark-blue end of the spectrum are most likely to be buried while those in the lime end are not. *Helix Propensity*: according to Jalview, residues in the purple end of the spectrum are most likely to occur in helices while those in the lime end are not.

Overall analysis of hydrophobicity throughout the protein suggests that as a whole the protein is mainly composed of hydrophilic residues sporadically interspersed with hydrophobic residues, with the exception being a short stretch of residues near the C-terminal regions of both DP motifs. Similar analysis of buried residues indicated that extremely few residues with a high propensity for remaining buried within the protein were found in the TPR Regions of the protein, while two short stretches of residues with a moderate buried index seemed to correspond to the same regions in the DP motifs

possessing a high hydrophobicity. Considering the structure of Hop as it is currently understood, and the fact that it is primarily cytosolic (Lassle, Blatch, Kundra, Takatori, & Zetter, 1997), it is no surprise that the most of the residues are low buried index residues with high helix propensity (Figure 2.3).

2.6.3 Phylogenetic Analysis

Several trees were constructed using several different tree-building methods and evolutionary models (see Appendix 3, Section A). The method that produced the most reliable tree, in terms of BIC scores (see Appendix 3, Section A, Table A3.1), bootstrap consensus (see Appendix 3, Section A, Figure A3.1), as well as consensus with other models, was the Maximum Likelihood method, using the rtREV evolutionary model. This protein tree was then compared to the Compara-Gene tree for PfHop on EnsEMBL.

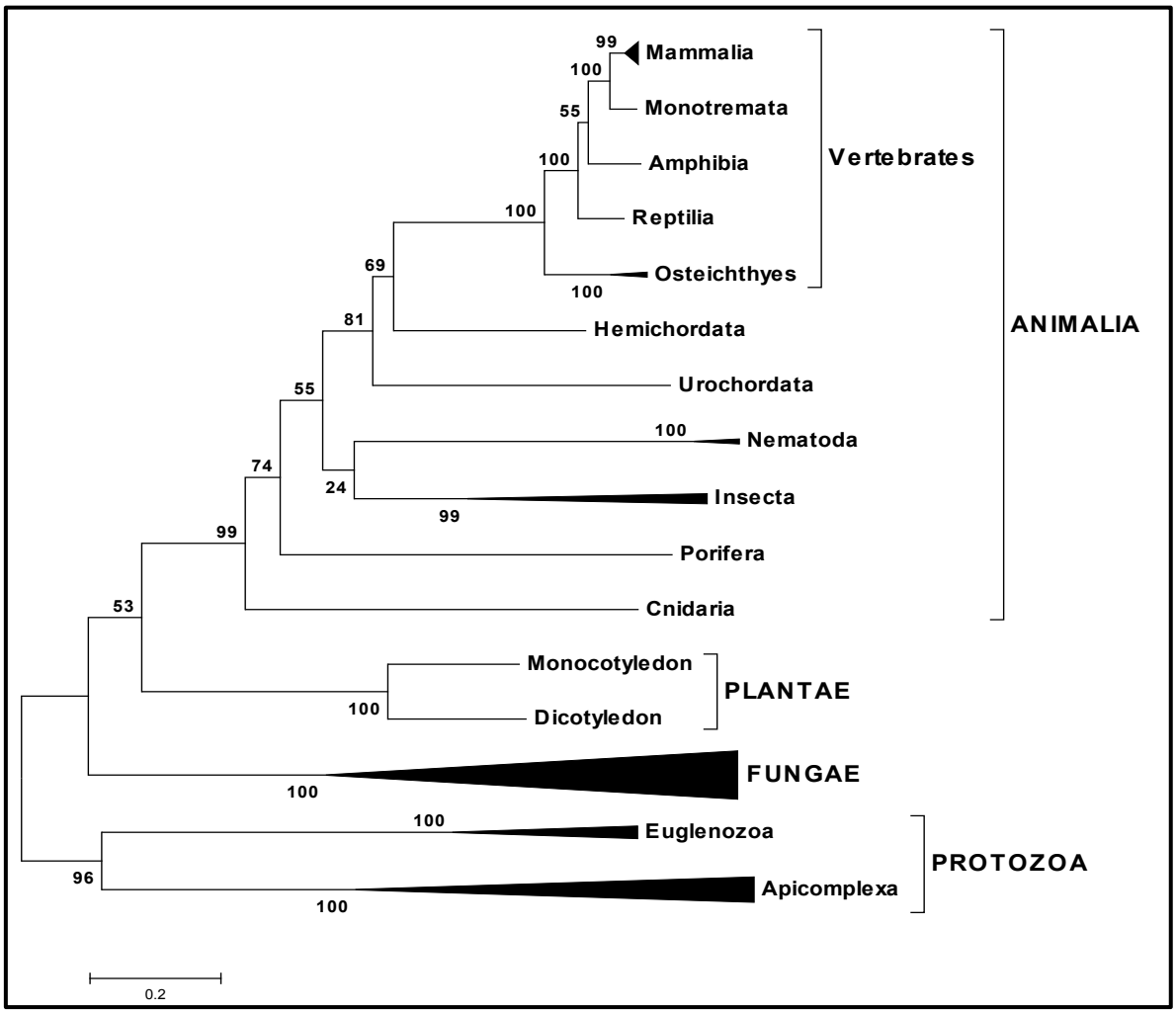


Figure 2.4: Major taxonomic groups in the rtREV Hop Protein Tree. (Graphic produced with MEGA's tree-viewer tool).

From Figure 2.4, it is plain to see that the statistical support values for the branching resolution of most the major groups are good (only five of fifteen major branches below 75, mostly within the plants and invertebrates). The branching resolution between the insect and nematode phyla has the lowest statistical support, with a bootstrap value of 24. This is the only region of disagreement between this tree and its bootstrapped support tree (see Figure A3.1, Appendix 3, Section A).

Comparison of Gene and Protein Trees

Analysis of the EnSEMBL Compara-Gene tree (see Appendix 4) for Hop indicates that the protein tree has some consensus (in terms of branching) with the gene tree. Both trees show that *O. anatinus* (Monotremata) is the first outgroup to the mammals. The gene tree indicates that the next outgroup is Reptilia (*A. carolensis*), followed by the Amphibia (*X. tropicalis*), whereas the protein tree indicates that the opposite is true. There is however poor statistical support for the branch that splits the amphibian from the mammals and monotreme (bootstrap value of 55) in the protein tree, which may help account for the disparity between the two trees. The two trees agree that the next outgroup is Osteichthyes (i.e. *D. rerio* and other ray-finned fish).

There is again disagreement as to the next outgroups to the vertebrates (i.e. the invertebrate animals); in the protein tree, the first outgroup is the Hemichordata (*S. kowalevskii*), followed by the Urochordata (*C. intestinalis*), then the grouped Insecta and Nematoda, followed by the Porifera (*A. queenslandica*) and finally, the Cnideria. In the gene tree there are no representatives for the Hemichordata and Cnideria, however, the poriferan is the first outgroup to the vertebrates, followed by a urochordate and finally the grouped nematodes and insects. Notably, there are four poorly supported areas of branching between the plants and invertebrates, making this area of the protein tree the least well resolved.

The gene tree indicates that the fungi are the first outgroup to the mammals, followed by the protozoans. Finally, Viridiplantae (Plantae) is outgrouped to all others. In the protein tree however, the plants are the next outgroup to the animals, followed by the fungi and finally the protozoa. This may be explained by poor resolution in branching between the plants and animals on the protein tree, or perhaps owing to a larger number of plant species used by the gene tree, thereby getting a more accurate estimation of variation within the

plants and thus allowing them to be more easily distinguished from the other groups. Owing to the discrepancy in methods and species used to build both trees, more in-depth analysis of branching is not really possible. Ideally, a gene tree should have been built using the genes for all 60 protein sequences for a more reliable comparison, but owing to time constraints this was not feasible.

As explained in the introduction, EnsEMBL Compara-Gene Trees are built with automated methods, and there are still many more automated trees on TreeFam than curated trees, and the gene tree for Hop should be treated with caution and used as a guideline for Hop phylogeny. The full protein tree for all Hop sequences used, indicating the various levels of taxonomic resolution used to assess the protein for later analysis, is displayed in high resolution on the next page for reference (Figure 2.5).

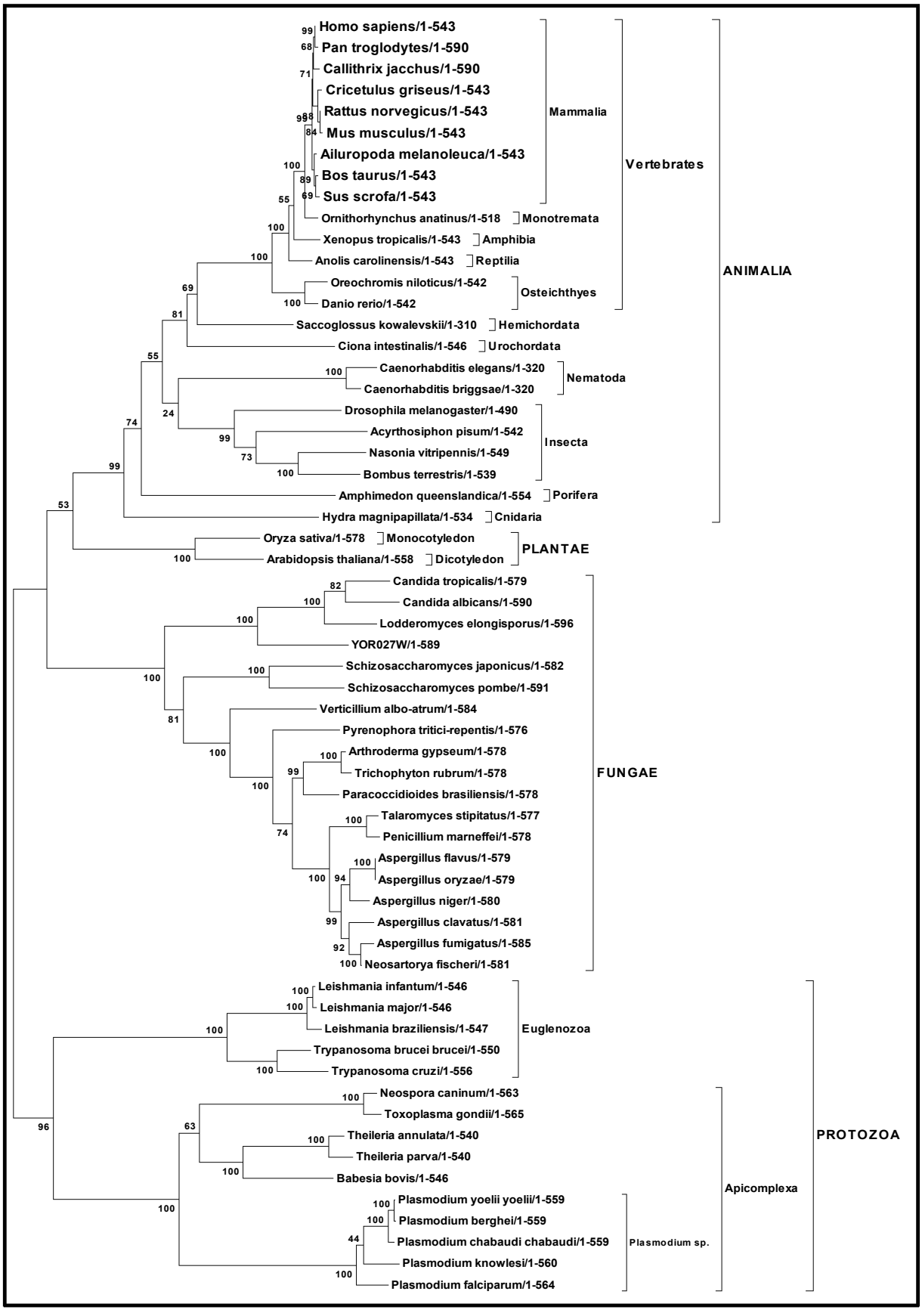


Figure 2.5: All taxonomic units in the rtREV Hop Protein Tree.

2.6.4 Domain Analysis

A maximum of 11 motifs (the longest being 113 residues), were found for all sequences of Hop, full-length protein (see Appendix 5, Section 3). The Mast output for all sequences (see Appendix 5, Section 3, Figure A5.3C) was reordered to reflect the top to bottom groupings of each species' Hop sequence in the protein tree for Hop (Figure 2.6).To the author's knowledge, this method of comparing MEME/MAST results according to the phylogeny of the sequences used is a novel approach. It yielded a satisfying view of domain sttructuring and conservation within selected taxonomic groups.

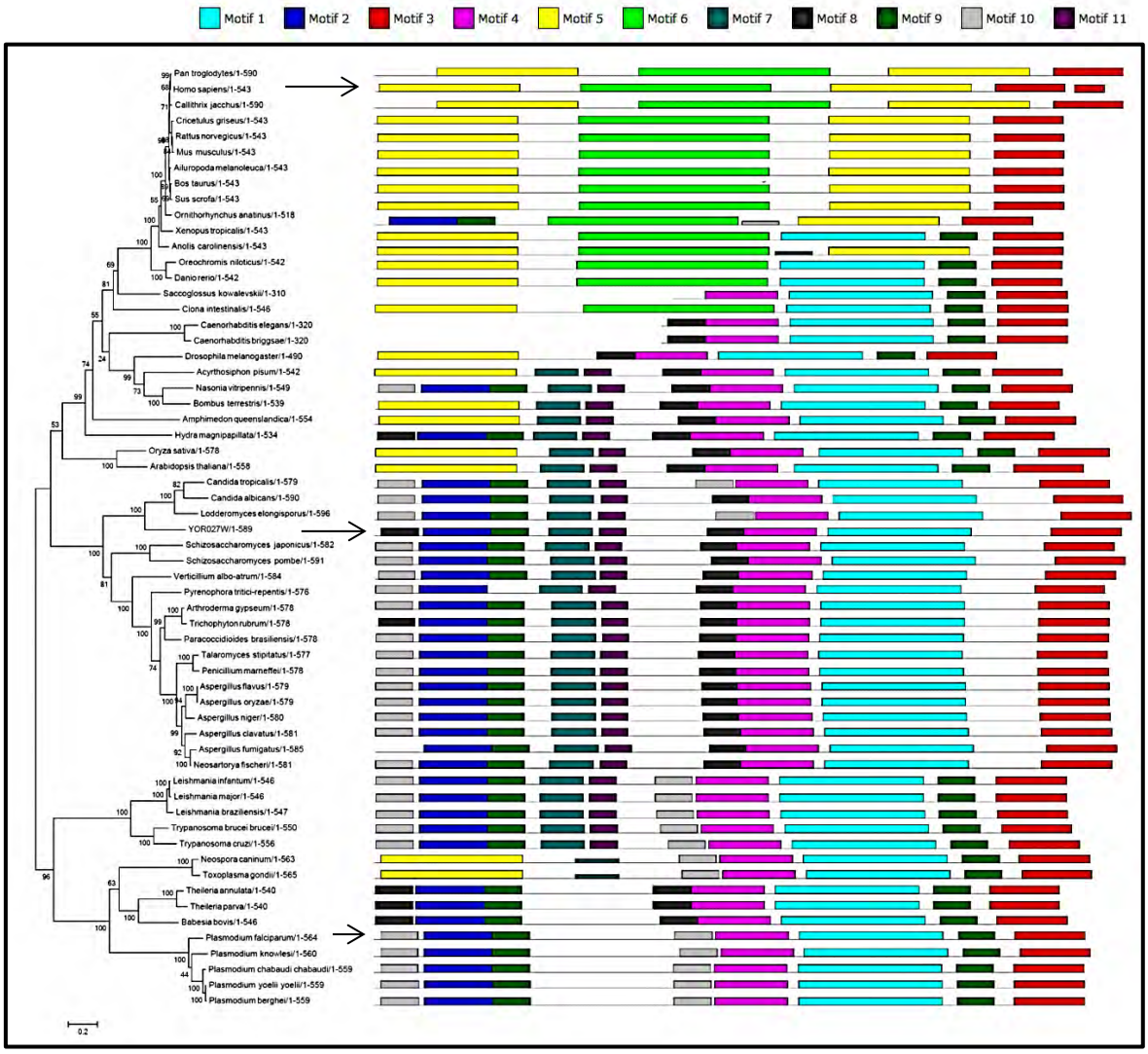


Figure 2.6: The Hop Protein (rtREV model) tree alongside Mast (Meme) results for each organism. See Appendix 5, Section 3, Figure A5.3.1-A5.3.11, for motif sequence logos). The three arrows are highlighting the (top to bottom) positions of HsHop, ScHop and PfHop, which correlate to Figures 2.14-2.16, respectively.

From Figure 2.6, one can see that there are various well-conserved domains that are reflected by motif representations. These motifs were aligned to condensed alignments of the various regions in Hop in Figures 2.7-2.12 and their significance is discussed on the following pages.

2.6.5 TPR1

For most of the species within the fungi and protozoan kingdoms, TPR1 is represented by Motifs 8 (dark grey) or 10 (light grey) at the N-terminal, 2 (dark blue) centrally and 9 (dark green) at the C-terminal, in that order. The exceptions are two Apicomplexan species *N. caninum* and *T. gondii*, and these species possess the same motif (motif 5, yellow) for TPR1 that the animal and plant Kingdoms share (Figure 2.6). Conversely there are several exceptions within the animal kingdom that share the same motifs for TPR1 with the fungi and protozoan kingdoms. One of these is a vertebrate, *O. anatinus* (commonly referred to as 'duck-billed platypus'). Considering how well Hop is conserved across the mammalian phylum and the fact that phylum Monotremata is the next outgroup to the mammals, the most likely reason the sequence for TPR1 is not represented by motif 5 is because the sequence is truncated at the N-terminal region (partial sequence, see Appendix 1, also see Figure 2.6), and it is this N-terminal region that allows motif 5 to be recognised. The other two animals are *N. vitripennis* (jewel wasp) and *H. magnipapillata* (cnidarian).

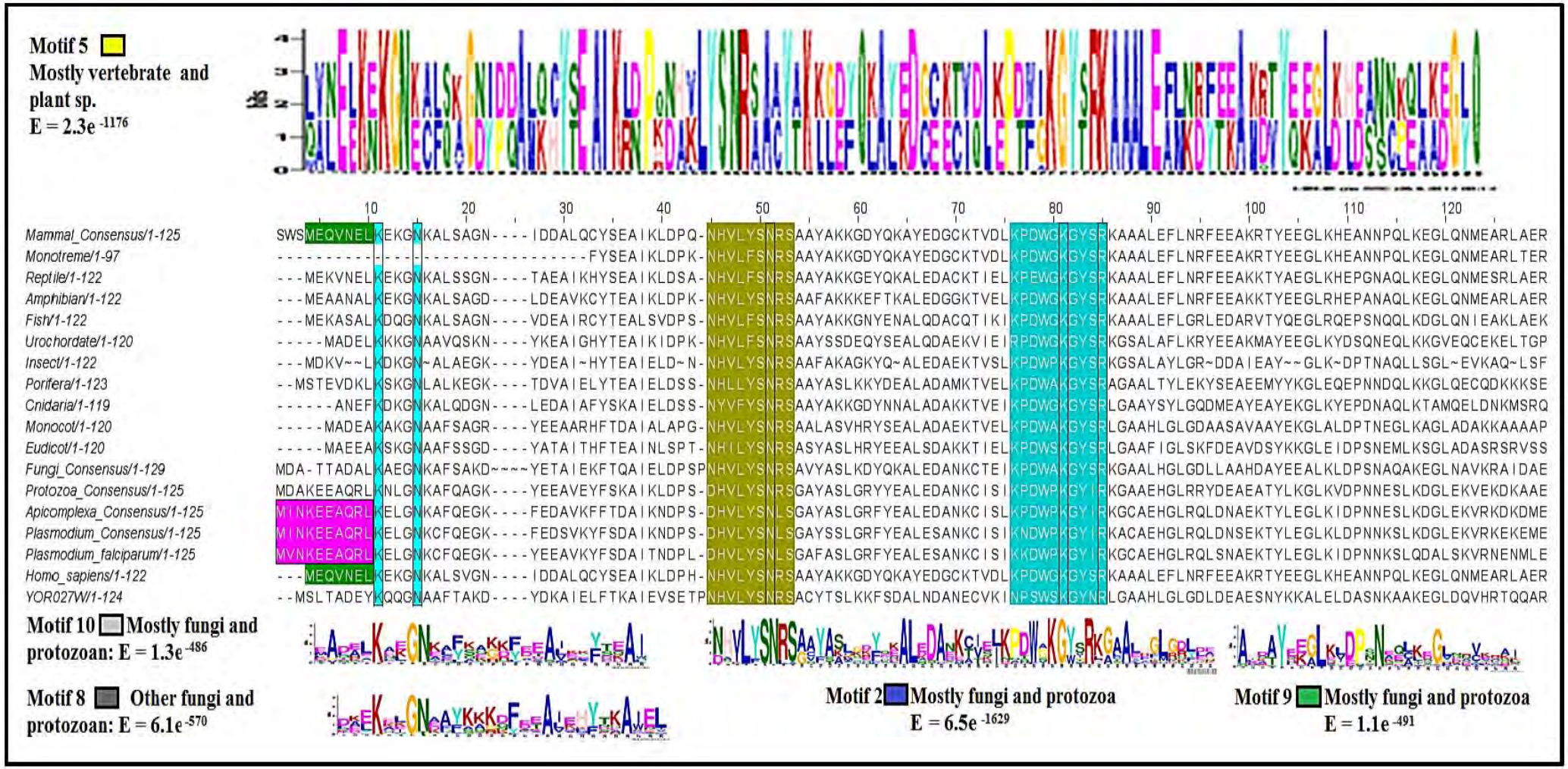


Figure 2.7: Condensed alignment representing the Hop TPR1 region for the major taxonomic groups (Figure 2.4) and species of interest, aligned with the relevant motifs numbered and coloured according to Figure 2.6. Several features are highlighted in the alignment (coloured regions) and the "carboxylate binding clamp" residues positions are bordered in black and coloured cyan. '~' indicates lack of consensus.

MAST analysis for TPR1 (see Appendix 5 Figure A5.3B) indicates that motif 5 shares high similarity to motif 2 (0.76) and motif 10 (0.63) which means that these two motifs may not be significantly different to the overlapping residue sites in motif 5. Motif 8 and 10 share the highest similarity (0.78), and are very likely not significantly different.

From Figure 2.7, one can see that the mammalian/vertebrate motif 5 (above the alignment) is very well conserved; almost all residues have a bit score greater than 3, while the fungal and protozoan motifs (displayed below the alignment) are more variable. Two interesting features of TPR1 that motif analysis did not detect are the start motifs; M[IV]NKEEAQL, well conserved for apicomplexans (pink box, Figure 2.7) and MEQVNEL which is well conserved for mammals (green box). Two other short motifs appear to be highly conserved across all species; NHVLYSNRS around position 50 and KPDWXKGYXR around position 80, and their conservation is reflected in motif 2.

2.6.6 DP1 and the Long Linker

As discussed in section 2.2.1, the DP1 and linker region is the least well aligned (see Figure 2.8). Interestingly, there are only two motifs that are specific to the DP1 region, and these are present only in certain invertebrates, plants, all fungi and the euglenazoans. These are motif 7 (teal) and 11 (mauve). The vertebrates and the urochordate possess motif 6 (lime green, Figure 2.6), which groups the hydrophobic region of DP1 and the long linker (Figure 2.8) as well as TPR2A (Figure 2.9). This could indicate that this whole region is a functional group in the vertebrates. MAST analysis indicates that Motif 11 has a high similarity score (0.65) to the overlapping residues on motif 6. This may indicate that the C-terminal region of DP1 is not significantly different in all species except the apicomplexans. The apicomplexa are definitely the exception to the rule for this region of the protein. They do not possess the DP/NP repeats that define the region (coloured by helix propensity, Figure 2.8), although they have chemically comparable substitutions in some of these DP positions (NP/NS).

Additionally, the apicomplexan long linker region is a glutamic acid-rich region, whereas for all other species this region is glutamic acid-rich, but interspersed with proline. In fact, in the vertebrates there is a short proline (x7) repeat (bordered in black, coloured in lime, Figure 2.11). This difference in the linker region may indicate a large structural difference for HsHop and PfHop, as the *P. falciparum* region most likely has alpha-helical secondary structure. Intriguingly, it may also provide clues for the alternate functioning of mammalian Hop in

neural cells. Several neurodegenerative disorders result from the cytotoxicity conferred by misfolded proteins. The cytotoxicity and aggregation property of a few mutant proteins are known to be modulated by the flanking sequences, such as a proline repeat tract (Siwach et al., 2011). The mammalian proline repeat tract has been shown to amend the cytotoxicity of a wide range of misfolded proteins coded by genetically engineered mutants. Additionally, the proline repeat tract may confer protection against the cytotoxicity of misfolded proteins by interfering with their conformation during translation and folding (Siwach et al., 2011). Hop's interaction with the Hsp90 chaperone suite places it in an ideal position to perform a similar function.

Proline is an exceptionally unusual amino acid in that it is cyclized, and is extremely restricted to certain phi and psi backbone conformations, thus a multi-proline peptide is even more limited. There is a repository of research reviewed by Williamson (1994) suggesting that a sequence of four or more proline residues in a row adopts a single preferred conformation in solution, known as the poly-proline II helix. This is an extended structure with three residues per turn. It is found primarily in some pancreatic polypeptide hormones and neuropeptides which mediate multi-protein complexes (Williamson, 1994).

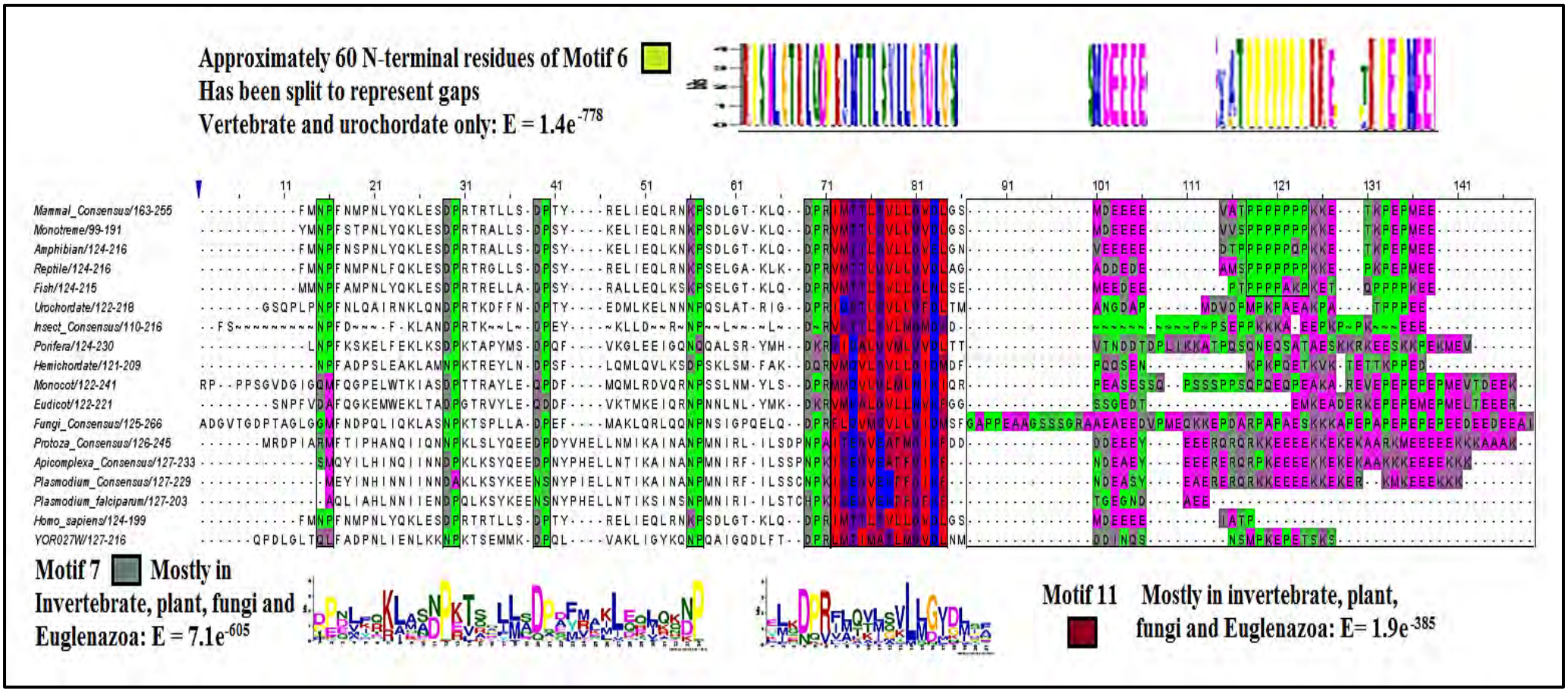


Figure 2.8: Condensed alignment representing the Hop DP1 region and long linker (position 86 – 148, coloured by helix propensity) for the major taxonomic groups (Figure 2.4) and species of interest, aligned with the relevant motifs numbered and coloured according to Figure 2.6. Several features are highlighted in the alignment; the DP repeats (coloured by helix propensity) and the C-terminal region of DP1 (position 72 – 84, coloured by hydrophobicity) residues positions are bordered in black. '~' indicates lack of consensus.

2.6.7 TPR2A

For almost the entire invertebrate, plant, fungi and protozoan species, TPR2A is represented by motifs 8 (dark grey) or 10 (light grey) at the N-terminal and motif 4 (bright pink) at the C-terminal (see Figure 2.4). The exception is the urochordate, *Ciona intestinalis* (marine seasquirt), and this species possesses the same motif (motif 6, green) for the C-terminal end of DP1, long linker and TPR2A that the vertebrate species share (Figure 2.6). MAST analysis (see Appendix 5 Figure A5.3B), indicates that motif 6 shares high similarity to motif 4 (0.66) and motif 8 (0.62) which means that these two motifs are not significantly different to the overlapping residue sites in motif 6, an overall indication that TPR2A may not be significantly different in all species analysed.

From Figure 2.10, one can see that the C-terminal portion of motif 6 (above the alignment) is very well conserved; almost all residues have a bit score greater than 3, while the fungal and protozoan motifs (displayed below the alignment) are more variable. Two interesting features of TPR2A that motif analysis did not detect in non-mammals are the first NLS, identified to be functional in mammals (Daniel et al., 2008; Odunuga et al., 2004).It also failed to distinguish a short, manually identified motif that appears to be highly conserved across most species; NXKAAVY (around position 51),its conservation is somewhat reflected in motif 4. Apart from the lysine at position 88, which displays arginine variability in the plants, all the "carboxylate binding-clamp" residues are highly (90-100%) conserved.

2.6.8 Linker Helix and TPR2B

For almost the entire vertebrate group, TPR2B is represented by the same motif that describes TPR1 in the vertebrates, invertebrates and plants, the N-terminal residues of motif 5 (yellow). The exceptions are the amphibian, *Xenopus tropicalis,* and the two ray-finned fishes, *Oreochromis niloticus* and *Danio rerio* (see Figure 2.6). These species are represented by the same motifs as those representing the invertebrate, plant, fungal and protozoan species (motif 1 in cyan at the N-terminal). All of these species that are non-fungal are further represented by the N-terminal residues of motif 9 (the same as that at the C-terminal of fungal and protozoan TPR1) at the C-terminal (see Figure 2.6). MAST analysis (see Appendix 5 Figure A5.3B) shows that none of these motifs have very high similarity to each other; an overall indication that TPR2B may be significantly different for groups possessing different motifs in this region.

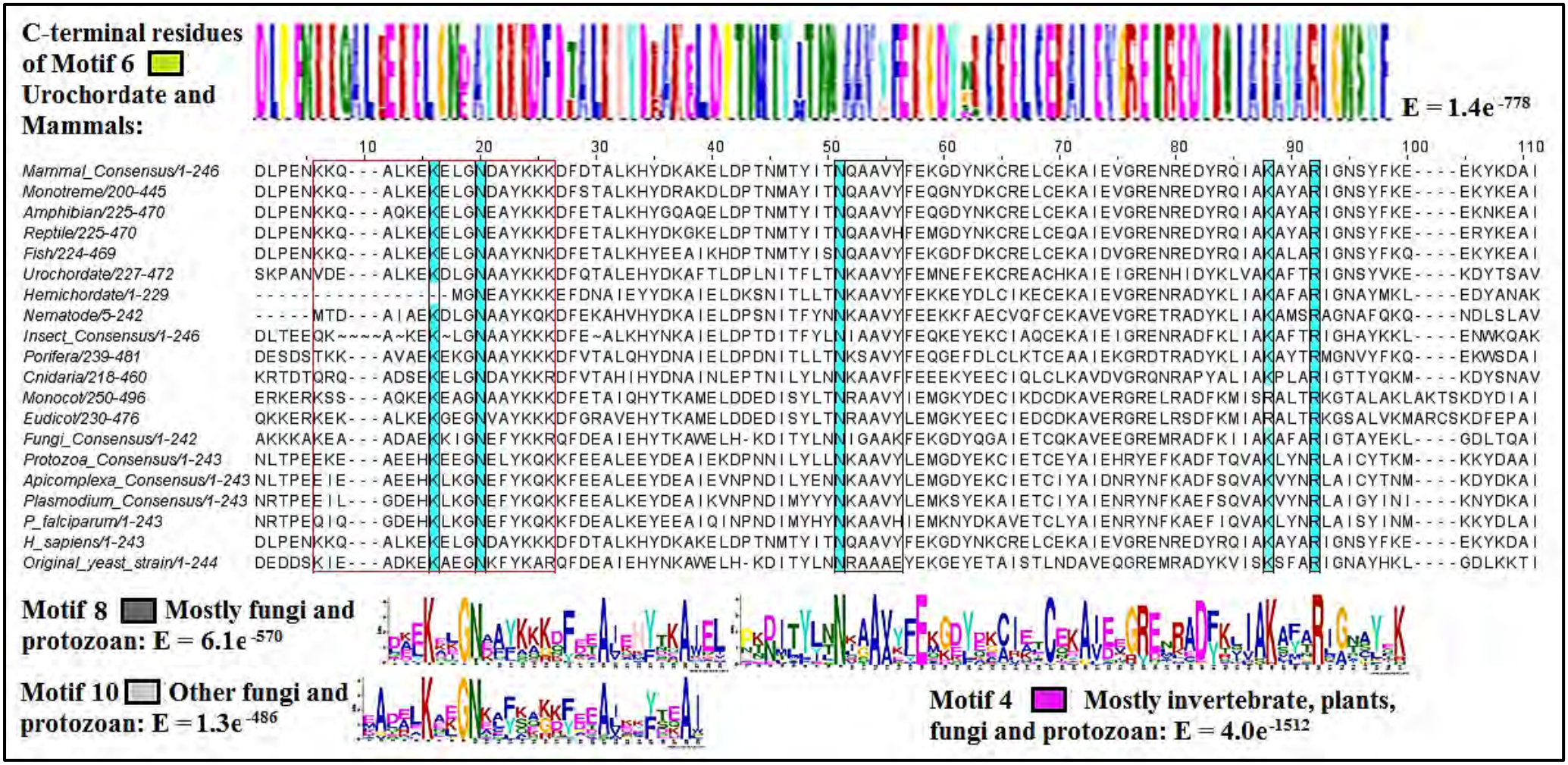


Figure 2.9: Condensed alignment representing the Hop TPR2A region for the major taxonomic groups (Figure 2.4) and species of interest, aligned with the relevant motifs numbered and coloured according to Figure 2.6. Several features are highlighted in the alignment; the first NLS (bordered in red) and the "carboxylate binding clamp" residues positions are bordered in black and coloured cyan. A short motif is highlighted from position 51 – 56. '~' indicates lack of consensus.

From Figure 2.10, one can see that the N-terminal portion of motif 5 (above the alignment) is very well conserved; almost all residues have a bit score greater than 3, while the invertebrate, fungal and protozoan motifs (displayed below the alignment) are more variable. Two interesting features of TPR2A that motif analysis did not distinguish are the second putative NLS (bordered in red), yet to be shown to be functional (Odunuga et al., 2004) and a short motif around position 190 that appears to be highly conserved across most species; LYSNRAA (bordered in black). Apart from the lysines at positions 154 and 220, which display arginine variability in the fungi, all the "carboxylate binding-clamp" residues (coloured in cyan and bordered in black) are highly (90-100%) conserved. The REY clamp residues (coloured in lime and bordered in black) are 100% conserved in all species.

In addition to the afore mention conserved regions, TPR2B has been reported to possesses the charged-Y motif; an 11-amino acid motif that was originally described for the non-concave Hsp90 interactions on the primary TPR motifs in FKB31 and FKB32 (Cheung-Flynn et al., 2003) and later identified as a feature of Hop (Odunuga et al., 2004). This motif has the consensus organization -+-+XϕYXXMFXXXX-, where - represents glutamic or aspartic acid, + represents lysine or arginine, ϕ represents a hydrophobic amino acid, and X represents any amino acid. It is interesting to note that the identifying tyrosine in this motif is the tyrosine found in the REY clamp. There is also a negatively charged amino acid five positions further downstream which may also be related to the functioning of this motif (Cheung-Flynn et al., 2003). The only motif describing this area bears some resemblance to this charged-Y motif; the overall analysis of the regex for this region in motif 1 is [+-][ϕ-]-+XϕYXXPEXXXX[ϕ-X]. This is not a distinct match so consensus sequences for the mammal and *Plasmodium* sequences were analysed separately in Figure 2.11.

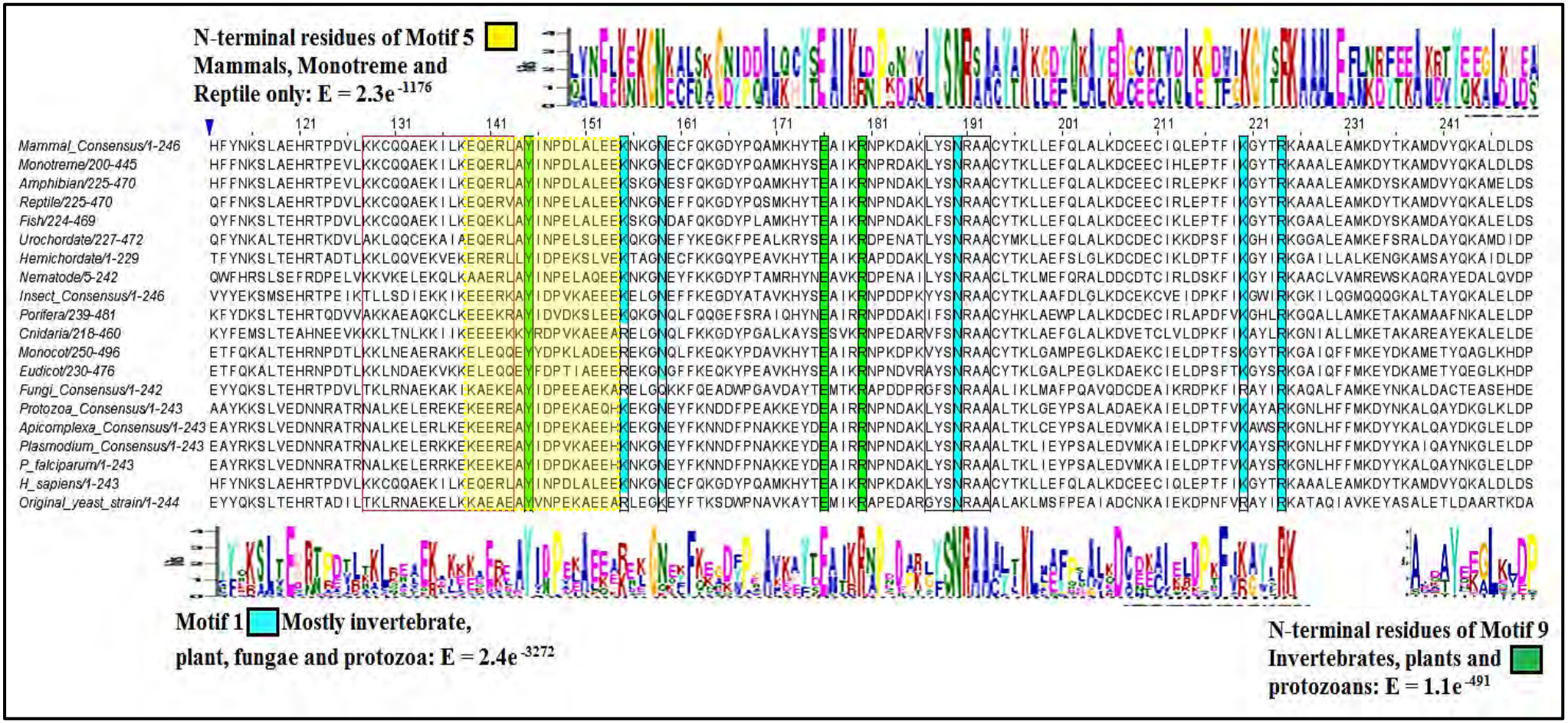


Figure 2.10: Condensed alignment representing the Hop TPR2B region for the major taxonomic groups (Figure 2.4) and species of interest, aligned with the relevant motifs numbered and coloured according to Figure 2.6. Several features are highlighted in the alignment; the second NLS (bordered in red), the charged-Y motif (shaded in yellow), the "carboxylate binding clamp" residues (bordered in black and coloured cyan) and the REY clamp residues (bordered in black and coloured lime). A short motif is highlighted from position 186 – 192. '~' indicates lack of consensus.

Charge Y REGEX:	-	+	-	+	X	ϕ	Y	X	X	M	F	X	X	X	X	-	
Mammalian_Consensus:	E	q	E	R	L	A	Y	I	N	p	d	L	A	I	E	E	
Plasmodium_Consensus:	k	e	E	R	E	A	Y	I	D	p	v	K	A	E	E	h	

Figure 2.11: Comparison of the mammalian and *Plasmodium* consensus sequences with the charged-Y motif consensus. Matches are in capitals, mismatches in small letters**.**

The mammalian consensus sequence for the charged-Y motif in Hop has six out of nine functional matches to the general consensus organisation for the charged-Y motif, while the *Plasmodium* consensus has only four out of nine matches. If this motif is truly a functional feature of Hop, it is a good starting point for assessing the alternate sites of Hop interaction with Hsp90, as this charged motif represents an Hsp90 interaction site on other proteins with TPR motifs.

2.6.9 Short Linker and DP2

The linker joining TPR2B and DP2 is relatively better aligned and more conserved than the long linker joining DP1 and TPR2A (bordered in black and coloured by helix propensity in Figure 2.12). It should be noted that a portion of this linker is represented by several of the C-terminal residues in motifs 5 and 9. As can be seen from the MEME results, DP2 is the only region in Hop that is represented by a single motif in all species; motif 3 (red) in Figure 2.6. This motif covers most of the C-terminal portion of DP2. Mast analysis (see Appendix 5 Figure A5.3B) shows that motif 3 has low similarity to all other motifs found, even those describing DP1: motif 7 (similarity = 0.49) and 11 (similarity = 0.51). This indicates that DP2 is distinct from DP1, in spite of the periodic DP repeats (coloured by helix propensity in Figure 2.12) and the hydrophobic stretch of C-terminal residues (coloured by hydrophobicity in Figure 2.12).

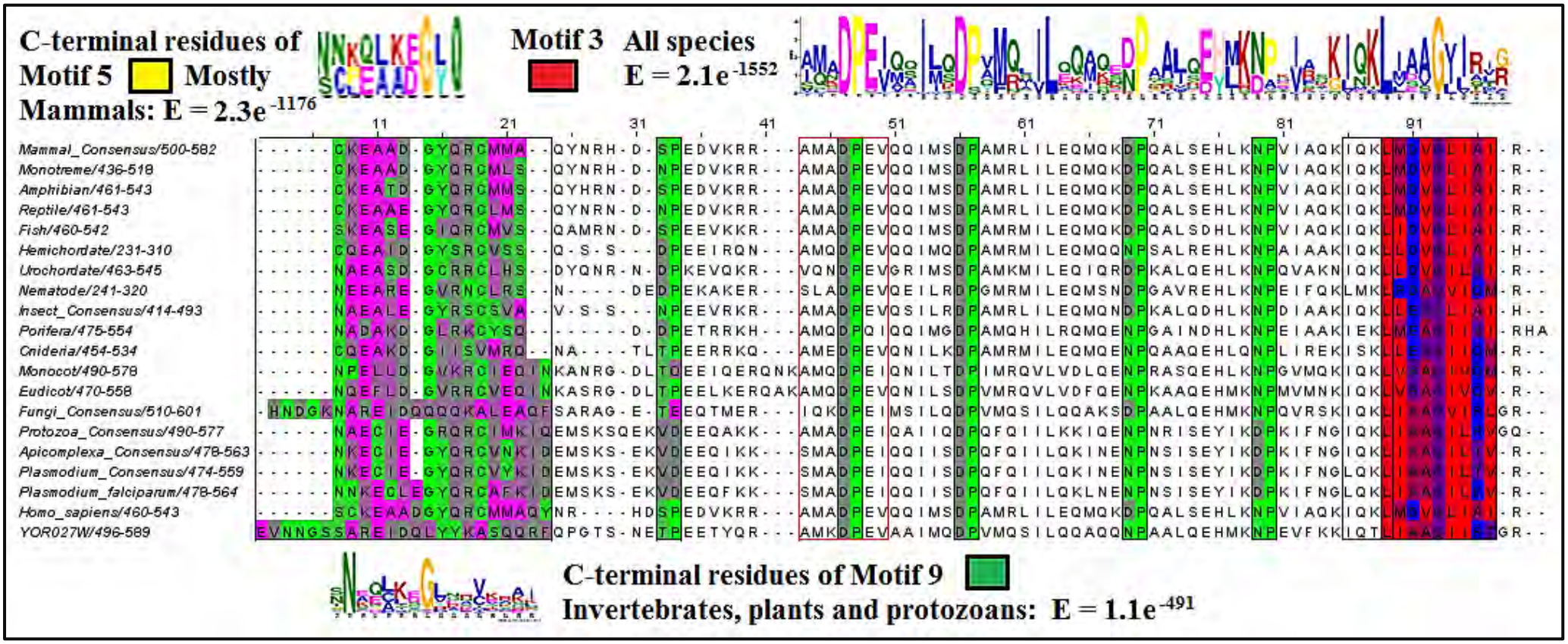


Figure 2.12: Condensed alignment representing the Hop DP2 region and short linker (position 1 – 24, coloured by helix propensity) for the major taxonomic groups (Figure 2.4) and species of interest, aligned with the relevant motifs numbered and coloured according to Figure 2.6. Several features are highlighted in the alignment; the DP repeats (coloured by helix propensity) and the C-terminal region of DP2 (position 89 – 96, coloured by hydrophobicity, bordered in black). Two short, manually identified motifs are highlighted from positions 44 – 50 and 86 – 89.

2.7 Conclusions

In an overview of Chapter two, one can conclude three main points that will assist with homology modelling and comparison of PfHop and HsHop. Firstly, Hop is very well conserved among all eukaryotes, particularly with regard to the "carboxylate binding-clamp" residues within the concave surfaces of the three TPR motifs, and in Hs- and PfHop these residues are identical for all three TPR motifs. Secondly, while Hop is very well conserved with regards to the TPR active site residues, domain organisation and other structurally important features such as the REY clamp, there are regions that are significantly different between human and *P. falciparum* and may be indicated by structural and functional disparities between the two proteins. These include what appears to be the significantly different (or possibly a complete absence of the) DP1 region in the Apicomplexa, the proline-rich stretch in the long linker region of the mammals and the difference in the overall structure of the "charged-Y" motif within TPR2B. These are the regions that are carefully analysed for structural and interaction studies and may provide sites that will one day prove to be exploitable for selective drug targeting.

Chapter 3: Homology Modelling

3.1 Introduction

Homology (or comparative) modelling is a strategy by which bioinformaticists try to bridge what is called the "sequence structure gap" (Eswar et al., 2008; Tastan Bishop et al., 2008). This gap is the result of an explosion of techniques that allows us to elicit gene and protein sequence information at ever-decreasing cost but with a corresponding lack of cheap and fast methods to determine three dimensional (3D) protein structures (Berman et al., 2003). Protein structure determination relies mainly on the time-consuming and expensive techniques such Nuclear Magnetic Resonance (NMR) spectroscopy, X-Ray Crystallography and High Resolution Electron Microscopy. There are also the theoretical methods of predicting the tertiary structure of a protein referred to as *ab initio* prediction or threading (Wu & Zhang, 2007).

The ability to accurately predict the tertiary structure of a protein, using nothing but the amino acid sequence with significant success has only ever been achieved with relatively small peptides. The computational requirements for modelling, taking into consideration all the physical and chemical attributes for all amino acid combinations, increases exponentially as sequence length increase. A relatively small protein would take several months to create a single correct model predicting its structure (Wu & Zhang, 2007). However, owing to the functional conservation of structure guided by natural selection and consistency of protein chemistry and folding governed by energy minima, it is feasible to model protein structures on closely related (homologous) template proteins via computational methods (Eswar et al., 2008; Tastan Bishop et al., 2008; Tastan Bishop & Kroon, 2011). This task relies heavily on complicated mathematical algorithms, requiring its automation. However, with machine governed automation there is always room for error; while computers are excellent at performing mathematical operations quickly, they cannot match human brains when it comes to discerning a functional pattern (such as protein fold) from massive amounts of data (Eiben et al., 2012; Khatib et al., 2011).

This means that model prediction is in fact a strategic process comprised of several steps. This includes identifying features of the target sequence that may assist with determining its functional structure. This can be done by identifying known functional domains through a

homolog search with BLAST, or using a secondary structure prediction program like JPRED. The next step is to find accurate template models based on crystal structure data. These templates should at the very least have 30% identity with the target (Vyas et al., 2012; Pei, 2008). However, in combination with MSA and structural alignment it is possible to build very accurate models even with sequence identities of less than 20% (Tastan Bishop & Kroon, 2011) One then uses a variety of programs to align the target sequence to the template and calculate a new model, which must be checked for aberrant loop regions that don't align and alternate rotamer conformations of side chains. If necessary, this model should then be refined or optimised (which is a simulation that allows atom positions in a model to vibrate into the lowest possible energy positions (Tastan Bishop & Kroon, 2011)). After refinement, several types of score can be used to test the accuracy of the model (Pawlowski, Gajda, Matlak, & Bujnicki, 2008). A human user has to guide the model prediction process by understanding both the concepts behind the software used and working around the limitations, double checking and back-tracking if necessary as a model is built (Tastan Bishop et al., 2008).

3.1.1 Modeller

Modeller is a program that comparatively predicts the tertiary structure of proteins (Sali & Blundell, 1993) utilising information on the intraprotein interactions within existing natively folded protein structures in the Protein Databank (PDB) (Berman et al., 2003; Berman et al., 2000). It is implemented exclusively within the Python programming environment. The PDB is a repository for 86 487 protein structures determined by methods such as NMR and X-ray crystallography. A subset of these structures was used to create a set of restraints that describe the range of spatial conformations a residue may occupy within a specific environment (i.e., solvent accessibility, contact, distance, torsional angle) (Sali & Blundell, 1993). These spatial restraints are expressed as a probability density function (pdf) for a specific structure or feature. Models are constructed to satisfy the spatial restraints for each fold or interaction feature based on alignment between the sequences of the template and the target (Eswar et al., 2008).

The 3D models are then optimised such that their overall pdf's violate the input (template) restraints as little as possible. Optimisation is an iterative process that involves alternating refinement steps; prediction of the rotamer orientations, recalculating the resulting shifts in the backbone atoms (for local energy minimization), then readjustment of the rotamers to the

new backbone, etc (Sali & Blundell, 1993). This will continue until the method converges at the global energy minima for the model. Slower refinement will theoretically result in more extensive sampling of the energy space for the model, and thus more likely result in lower energy (more native-like) models (Misura & Baker, 2005).

Usually, only a single, good quality template (>30% sequence similarity, greater target coverage, etc.) is required for model building. However, several templates may need to be used to account for large gapped regions, or low sequence similarity. If the target-template alignment contains several different templates with many insertions and/or deletions, it is necessary to calculate multiple models for the same alignment (Eramian et al., 2008; Eswar et al., 2008; Tastan Bishop et al., 2008). This allows for better sampling of the different template segments and the conformations of the unaligned regions, and will usually result in a more accurate model, through selection of the best model (usually that with the lowest energy) based on a comparative score, such as the DOPE score (Shen & Sali, 2006). The most accurate models usually have Cα RMSD values within 0.5Å of the true native structure (Eswar et al., 2008). However, for targets aligned to template/s with high sequence and structural similarity and few, small, gapped regions (less than five residues), building multiple models will not necessarily result in better accuracy of the best model produced.

3.1.2 PyRosetta

PyRosetta is a stand-alone implementation of the Rosetta molecular modelling package (Leaver-Fay et al., 2011) that can be installed locally and run within a simple programming environment (Chaudhury, Lyskov, & Gray, 2010). Having similar computational performance to Rosetta, it can be employed for functions such as protein docking, protein folding, loop modelling and design using the major Rosetta sampling and scoring functions (Lyskov & Gray, 2008; Misura & Baker, 2005; Tyka et al., 2011). This is possible because it possesses Python bindings to libraries and databases that constitute Rosetta functions. The program may be used in two ways; script-based and interactively, using a customised Python shell called iPython. The interactive shell contains a number of help features; autocomplete commands for familiarising one's self with the software and the ability to link the command shell to PyMOL for real-time visualisation of molecular changes during building, docking, refinement and scoring (Baugh, Lyskov, Weitzner, & Gray, 2011). Customized scripting is best suited to developing reusable packages and tools for research (Chaudhury et al., 2010). See, for

example, the “Tkinter Minimisation Toolkit” produced by Jared Adolf-Bryfogle and accessed at http://www.rosettacommons.org/node/2344 (Fox Chase Cancer Center, Drexel College of Medicine).

3.1.3 Refinement in Rosetta

The ClassicRelax mode in Rosetta carries out the task of structural refinement (Tyka et al., 2011). After creating a crude model generated by *ab-initio* structure prediction or a low quality model created with homology modelling, one utilises the Rosetta energy function to refine the structure while searching through conformational space. This is a movement of the protein backbone and side-chain torsion angles by relatively small amounts from the starting structure. It has been shown to dramatically lower the full-atom energy of a model as it uses the Rosetta energy function, which is highly sensitive to steric clashes, to improve side-chain interactions significantly (Khatib et al., 2011; Misura & Baker, 2005; Tyka et al., 2011). A more flexible, modern version of the initial relax algorithm is FastRelax; it functions by running many alternating side chain repacking and minimisation cycles of non-deterministic Monte Carlo simulated annealing, which randomly searches combinations of side-chain conformations chosen from a library of possible rotamers. The structure’s RMSD can oscillate up to 2-3Å from the starting conformation during these minimisation cycles (Leaver-Fay et al., 2011; Tyka et al., 2011).

Relax does not perform extensive refinement and only searches the immediate local conformational space (Tyka et al., 2011). Researchers can create and customize their own minimisation protocols with this capability through the creation of “mover” methods and “packer” tasks in Rosetta. These tasks and methods can be restricted to certain aspects of the model in question, such as repacking of side chains only, or minimisations of loop regions only (Chaudhury et al., 2010; Leaver-Fay et al., 2011; Tyka et al., 2011).

The ClassicRelax protocol in Rosetta has been validated through an application to the refinement of several *de novo* structures (Misura & Baker, 2005). Analysis showed that there were important structural differences between the refined models and the refined idealized native structures owing to errors in local backbone and side-chain conformation, as well as strand alignment (Misura & Baker, 2005). This study suggested that problems with backbone and side chain conformation are common when the true structure contains an energetically

unusual feature, such as an energetically unfavourable side chain rotamer or local backbone conformation that results in steric clashes. Additionally, low energy models with an excess of helical secondary structure relative to the true native structures are biased for by the Rosetta energy score (Lazaridis & Karplus, 1998). This is because the rotamer conformation space for well-formed helices is condensed and more thoroughly sampled. Additionally, helices contribute more favourable attractive energies and have stable hydrogen bonding networks, contributing favourably to the overall hydrogen bond energy. A much larger rotamer conformation space must be sampled to align and pair neighbouring beta strands and result in comparable energetically favourable structures and overall Rosetta energies (Misura & Baker, 2005). This phenomenon may favour the minimisation of Hop homology models as it is comprised entirely of helices and loop orlinker secondary structures. Conversely, should the native structure for PfHop differ from homologs used as templates by containing beta strands or sheets, this step may introduce errors.

3.1.4 Structure Quality Validation

Crystal structure quality is primarily assessed through resolution, R-value and R-free scores. Resolution is usually recognised as a measure of the level of detail present in the diffraction pattern in an X-ray crystallography experiment and the resulting electron density map. The resolution of a model can be anything from 1-2 Å, which is highly ordered, making it easy to accurately predict the position of every atom in the electron density map, and 3 Å or higher, which is disordered; showing only the basic surface of the protein chain. There are other methods, where understanding of the atomic position is not as important as getting an overview of tertiary and quaternary structure. High-resolution TEM techniques, such as cryo-EM, would result in low resolution images which may be used to create refined models with resolution as low as 15 Å using computational techniques (Southworth & Agard, 2011). Electron cryo-microscopy, a recently developed TEM technique, has been utilised to create images of virus particles at a resolution of 3 – 4 Å, almost atomic resolution (Grigorieff & Harrison 2011).

The atomic model built during the process of determination through X-ray crystallography is used to calculate a simulated diffraction pattern based on that model. The R-value indicates the closeness of the match between the simulated diffraction pattern and the experimentally-observed diffraction pattern. An R-value of 0 indicates a perfect match, while 0.63 is the score

one would expect from a random selection of atoms. The values for most structures in the PDB are around 0.20. The problem with relying solely on R-values is that subsequent refinement is often used to improve the atomic model and make it better fit the experimental data (improve the R-value), which introduces bias. To assess the extent of this bias, a second indicator is calculated; the R-free value. Before refinement begins, approximately 10% of the observed data are removed from the data set, leaving only 90% of the data free for use during refinement. The R-free value is then calculated by seeing how well the refined model predicts the 10% control data. Ideally the R-free will be similar to the R-value, in practice however, it is slightly higher; with a value of about 0.26.

Traditionally, most model quality assessment programs focus on the evaluation of the entire protein structure using some form of qualitative or relative scoring function (Chaudhury et al., 2010; Shen & Sali, 2006; Zemla et al., 2002) rather than detection of correct and incorrect regions. The most accurate models generally have the lowest free energy of all empirically predicted models under physiological conditions (Sali & Blundell, 1993; Shen & Sali, 2006). Free energy functions enable the prediction and assessment of the best models from a subset, where the free energy surface of a protein can be derived by thoroughly sampling the potential energy surface defined by a molecular mechanics force field, such as those utilised by CHARMM (Eramian et al., 2008). Owing to inherent errors in potential energy functions and long computational run time, an alternative to calculating the free energy surface of a protein is to use a scoring function whose global minimum corresponds to the true structure from a sample of random structures of different sequences deposited in the Protein Data Bank (Eramian et al., 2008; Shen & Sali, 2006).

Such a scoring term is referred to as a “knowledge-based” or “statistical” potential and is not an explicit physical energy term, but one that may be used qualitatively. Statistical potentials are grouped by several characteristics: protein aspect representation (e.g., residue centroids, Cα-atoms and all-atoms), the restrained spatial features, and the reference state (Lazaridis & Karplus, 1998, 1999). Scores utilising all-atom representation, such as the DOPE and Rosetta energy scores, are generally more accurate than those for an amino acid residue representation. This usually results in increased accuracy for best model selection.

Programs that are capable of evaluating specific regions or residues of the model in question often recommend evaluating a score that is averaged over a long stretch of residues. Two such

programs are ANOLEA and VERIFY3D. ANOLEA uses a highly sensitive atomic mean force potential (AMFP) to calculate the non-local energy profile of a structure (in graphical output). Very high scores represent areas in the model where stereo-chemistry is not feasible and misalignments have occurred. These regions almost invariably point to loops or areas where alternate rotamer conformations are possible (Melo & Feytmans, 1998). VERIFY3D was developed to produce a 3D profile of the atomic coordinates of a structure which, when the model is accurate, matches that of its own sequence resulting in high scores (Luthy, Bowie & Eisenberg, 1992).

3.1.5 Normalised DOPE Score

The Discrete Optimised Protein Energy (DOPE) score is an atomic distance-dependent statistical potential used to evaluate structures. It is based on a physical reference state that corresponds to non-interacting atoms within a uniform sphere, with the finite size and spherical shape of proteins dependent on a sample native structure (Shen & Sali, 2006). The normalized version of the DOPE score (N-DOPE, Eramian et al., 2008) is commonly used; it is a standard Z-score derived from basic statistics of raw DOPE scores (mean and standard deviation); where positive scores are likely to be poor models, while scores lower than -1 are likely to be closest to the native structure (Eswar et al., 2008). The performance of normalised DOPE score was superior compared to thirteen other commonly used statistical potentials, however, it increased when the overall accuracy of the subset of models being assessed was increased (Eramian et al., 2008). This is a commonly observed trend for other scoring functions, if to a lesser extent. Additionally, the average correlation of normalised DOPE to Rosetta and RMSD scores exhibits slight improvement when limited to high-accuracy targets (Tastan Bishop & Kroon, 2011).

3.1.6 Rosetta Energy Score

The Rosetta energy score was initially developed to distinguish misfolded models from native structures (Lazaridis & Karplus, 1998). A revised function for proteins in solution utilizes an improved CHARMM19 polar hydrogen potential energy function complemented by a simple Gaussian model for the solvation free energy and described in detail by Lazaridis and Karplus (1999). Although many aspects of this scoring function are heuristic, it differs significantly from ‘‘knowledge-based’’ potentials and can be used for molecular dynamics simulations as

well as for comparing a high-accuracy subset of structures (Lazaridis & Karplus, 1999). The Rosetta energy function involves significant approximations, such as modelling solvent implicitly rather than explicitly and neglecting long-range electrostatics which results in assignment of higher energies to the more accurate models than to the incorrect models. These cases appear to be primarily for native structures that are not globular (Lazaridis & Karplus, 1998), like Hop, or that have differences in helix and sheet content between the native structure and the unrefined models for reasons explained previously. However, the Rosetta energy performs relatively well regardless (Eramian et al., 2008) and when considered alongside the normalised DOPE, has a good correlation to the DOPE score for high-accuracy models (Tastan Bishop & Kroon, 2011).

3.1.7 MetaMQAPII

MetaMQAPII is a program designed to accurately assess the local structural quality of a model as well as overall tertiary accuracy (Pawlowski et al., 2008). This program is a meta-server that incorporates results from eight other model quality assessment servers; VERIFY3D, ProSA (Wiederstein & Sippl, 2007), ANOLEA, BALA-SNAPP (Krishnamoorthy & Tropsha, 2003), TUNE (Lin, May & Taylor, 2002), REFINER (Boniecki et al., 2003) and PROQRES (Wallner & Elofsson, 2006). MetaMQAPII assesses each residue in a structure by first placing it into one of 315 electrostatic environment groups. For each of these groups, a unique linear regression model has been developed to determine the RMSD of a residue within that group from its location in the native structure, depending on how well it was scored by the combination of eight different quality assessment servers (Pawlowski et al., 2008). The meta-server then outputs a PDB coordinate file of the model, where the B-factors for each residue have been replaced with the ranking score from the overall assessment. These scores can then be visualised in most molecular visualisation software and are represented on a colour spectrum from blue (correct, low RMSD from native) to red (incorrect, high RMSD from native) to quickly identify problematic regions (Pawlowski et al., 2008). Along with this coordinate file, MetaMQAPII also returns a log file containing the predicted GDT_TS score (Zemla et al., 2002), an overall RMSD value describing the predicted deviation in angstroms from the true protein structure, and a table containing all residue scores returned from each of the eight model quality assessment programs used by the meta-server.

3.2 Methods and Software

3.2.1 Structure Retrieval

All structures used for homology modelling were retrieved from the PDB Repository. Except for 4GCO, all structures were referenced structures associated with peer-reviewed literature. The structure 4GCO from *C. elegans* is not associated with a publication; more information can be found from: http://kiemlicz.med.virginia.edu/mcsg/deposits/index. The structures were viewed with visualization programs such as PyMOL (DeLano, 2002), Discovery Studio Visualizer (Accelrys Software Inc., 2007) and MATLAB's MolViewer (The MathWorks Inc., 2009).

3.2.2 Homology Modelling

Non-protein atoms (water and ions) were removed from coordinate files by selecting and saving only the peptide chains, in PyMOL. Homology models were generated using Modeller (Eswar et al., 2008; Shen & Sali, 2006) in high throughput fashion to predict PfHopTPR2A-PfHsp90 (C-terminal motif) and PfHopTPR2B (and TPR1)-PfHsp70 (C-terminal motif) complexes using published templates as well as HsHop (relevant TPR regions) in complex with the PfHsp90 and PfHsp70-x C-terminal motifs. For every template-target combination, 100 models were built using the standard "automodel" routine of the program with a very slow refinement option. See Appendix 2, Section B, for examples of modelling scripts used.

3.2.3 Model Validation

The quality of the homology models was evaluated by calculating several parameters. While the primary indicator used to select for top models was the N-DOPE Z score, two other energy scores; the Cα-RMSD (from template) and Rosetta Energy scores were calculated for all models. See Appendix 6, Section C, for examples of the scripts used to calculate and compare these various scores. Ramachandran plots produced in Rampage (Lovell et al., 2003), as well as visualisation of MetaMQAPIIscores obtained for each monomer from the MetaMQAPII web server (Pawlowski et al., 2008) were used to assess model quality at the level of individual residues.

3.2.4 Post-modelling Optimisation and Modification

Models were further refined using the Tkinter Minimisation Toolkit (produced by Jared Adolf-Bryfogle, Lab of Dr. Roland Dunbrack, Fox Chase Cancer Center, Drexel College of Medicine) which utilises the Classic- and FastRelax protocols in PyRosetta.

3.3 Results and Discussion

3.3.1 Template Analysis Summary

This section contains a final summary of all templates used to create working human and *P. falciparum* homology models. The data pertaining to the templates summarised in Tables 3.1 and 3.2 will be discussed in depth in sections in Section 3.3.2.

Table 3.1: Complex structure template summary.

Template	1ELW	3UQ3	3UPV	SchmidCYS
Organism	*H. sapiens*	*S.cerevisiae*	*S. cerevisiae*	*S. cerevisiae*
Hop Domain	Chain A: HsTPR1	Chain A: ScTPR2A&B	Chain A: ScTPR2B	Chain B: ScTPR2A&B (chain B is 3UQ3)
Hop Partner/s	Chain C: HsHsp70-GPTIEEVD C-terminal motif	Chain B: ScHsp90-MEEVD C-terminal motif Chain C: ScHsp70-EVD C-terminal motif	Chain B: ScHsp70-PTVEEVD C-terminal motif	Chain A: ScHsp90 M and C Domains
Structure Determination	XRD	XRD	XRD	Combination of Spin-labelling and Docking Techniques
Resolution	1.60 Å	2.60 Å	1.60 Å	N/A
R-Value	0.180	0.222	0.186	N/A
R-Free	0.215	0.279	0.254	N/A
N-DOPE Z	-2.569	-1.562	-2.352	-1.069
Rosetta Energy	-257.680	-229.453	-161.600	199.047
Minimised N-DOPE Z	-2.536	-1.666	-2.378	-1.166
Minimised Rosetta Energy	-368.205	-811.139	-400.139	-1680.626
Interaction Energy*	-10.709	Overall: -21.493 2A: -13.434 2B: -8.058	-9.96	5.566
Binding Energy*	42.112	Overall: 298.502 2A: 298.874 2B: 303.651	110.386	-1404.36
% ID to corresponding domain in target	PfHopTPR1: 39.84%	HsHopTPR2AB: 46.36% PfHopTPR2AB: 36.63%	PfHopTPR2B: 49.61% HsHopTPR2B: 45.61%	HsHsp90: 61.00% PfHsp90: 36.00%
Reference	(Scheufler et al., 2000)	(Schmid et al., 2012)	(Schmid et al., 2012)	(Schmid et al., 2012)

*In terms of Rosetta energy scores. See Chapter 4 for explanation of how the binding and interaction energies were calculated.

Table 3.2: Single structure template summary.

Template	2LLV	2LLW	4GCO
Organism	*S. cerevisiae*	*S. cerevisiae*	*C. elegans*
Hop Domain	Chain A: ScDP1	Chain A: ScDP2	Chain A: CeTPR2B
Structure Determination	Solution NMR Average of 21 Models	Solution NMR Average of 21 Models	XRD
Resolution	N/A	N/A	1.60 Å
R-Value	N/A	N/A	0.179
R-Free	N/A	N/A	0.219
Modeller's N-DOPE Z	-1.680	2.405	-2.168
Rosetta Energy	182.126	24781.403	-95.861
Minimised N-DOPE Z	-2.279	-0.398	-2.189
Minimised Rosetta Energy	-140.029	4137.759	-352.763
% ID to corresponding domain in target	HsHopDP1: 31.17% PfHopDP1: 24.32%	HsHopDP2: 40.23% PfHopDP2: 33.33%	PfHopTPR2B: 49.11%
Reference	(Schmid et al., 2012)	(Schmid et al., 2012)	(Osipiuk et al., 2012)

3.3.2 Template Analysis for TPR structures

Table 3.3: Percentage identity for various domains to be modelled from yeast templates in human and *P.falciparum.*

Yeast Template	Human	*P. falciparum*	Human and *P.falciparum*
TPR1	34.68%	36.00%	39.84%
DP1	31.17%	24.32%	22.64%
TPR2	46.36%	36.63%	42.69%
DP2	40.23%	33.33%	36.78%

3.3.2.1 Template for Modelling TPR1

The first functional domain of Hop is the TPR1 region. A single, high quality structure for this region (in human Hop) in complex with HsHsp70-GPTIEEVD C-terminal motif was used to model this region in PfHop (Scheufler et al., 2000). This crystal structure has excellent resolution (1.6Å) with low overall R-value and R-free scores.

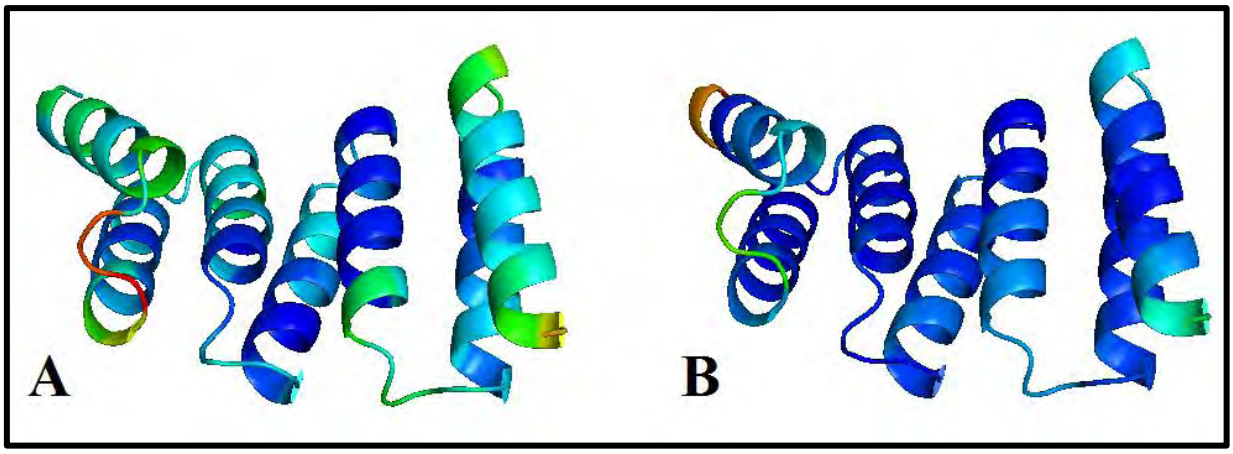


Figure 3.1: A) MetaMQAPII rendition of chain A 1ELW B) MetaMQAPII rendition of minimised 1ELW chain A.

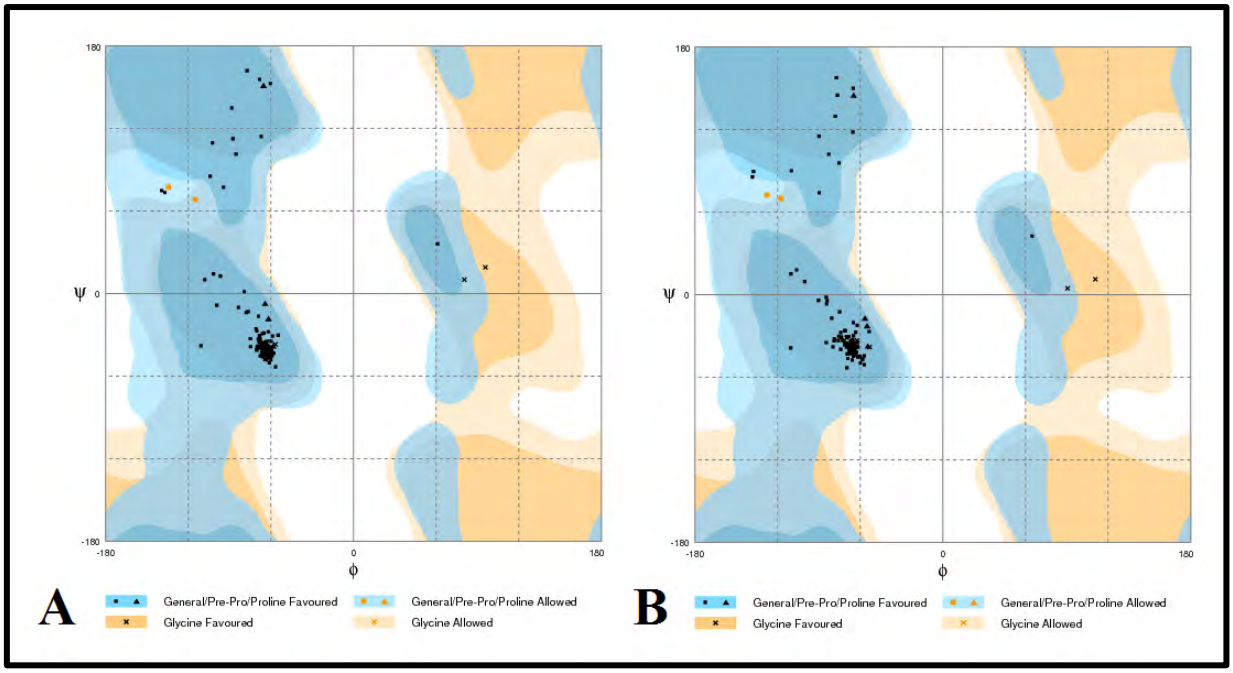


Figure 3.2: Ramachandran plots for A) 1ELW and B) minimised 1ELW. Black triangles and squares represent amino acids in the "Favoured" regions, Orange. Triangles and squares represent amino acids in the "allowed" regions.

From Table 3.1, and Figure 3.1, it was seen that the overall beta-factor scores as well as the Rosetta energy for the minimised version of 1ELW were lower. However, the original version of the template still had the lowest N-DOPE Z score and Ramachandran plots (Figure 3.2) indicated that the original and the minimised structures were of equal quality.

3.3.2.2 Templates for Modelling HopTPR2A&B in Complex with C-terminal Partner Peptides

The third and fourth functional domains of Hop are the TPR2AB regions. Two high quality structures for this region (in yeast Hop) in complex with ScHsp90-MEEVD and ScHsp70-PTVEEVD C-terminal motifs were used to model this region in both *P. falciparum* and human Hop (Schmid et al., 2012). 3UQ3 is a relatively good quality structure (resolution of 2.6Å) derived through XRD of ScHopTPR2A&B in complex with 5 residues of ScHsp90 C-terminal motif (MEEVD) and three residues of ScHsp70 C-terminal motif (EVD).

While the well characterised "double carboxylate binding clamp" appears to interact primarily with the last four residues of partner protein i.e. EEVD, the interactions described above cannot discriminate between the C-terminal motifs of Hsp70 and Hsp90 and additional contacts are made with residues upstream of the EEVD motif (Scheufler et al., 2000). Previous studies (in humans) have shown that these contacts are important; the C-terminal heptamer motif (GPTIEEVD) of Hsc70 will bind to TPR1 with the same affinity as the complete C-terminal domain of Hsp70/Hsc70, but the C-terminal tetramer motif (EEVD) resulted in a sharp drop in the affinity of TPR1 for Hsp70. Additionally an extension of this motif (IEEVD) to match the length of the Hsp90 (MEEVD) peptide, still bound with significantly weaker affinity than the heptamer peptide (Scheufler et al., 2000).

Similar studies in yeast published earlier in 2012 show that the TPR2A domain possesses a hydrophobic pocket to accommodate the methionine in MEEVD of Hsp90 as observed in the structure of 3UQ3. TPR2B lacks this pocket but contains a selective binding cavity for the threonine in PTVEEVD in Hsp70, although, the peptide backbone has to adopt an energetically unfavourable helical turn to bind at this position. As TPR1 is so similar to TPR2B in humans (as established in Chapter 2), it is logical to assume that the same improved interaction with longer C-terminal peptide is true for TPR2B. For this reason the additional template, 3UPV, representing the yeast TPR2B region in complex with a longer Hsp70 C-terminal peptide (PTVEEVD) than that available in 3UQ3 (EVD), was used in combination with 3UQ3 to create homology models of human and *P. falciparum* HopTPR2A&B with longer C-terminal Hsc70 and Hsp70 motifs.

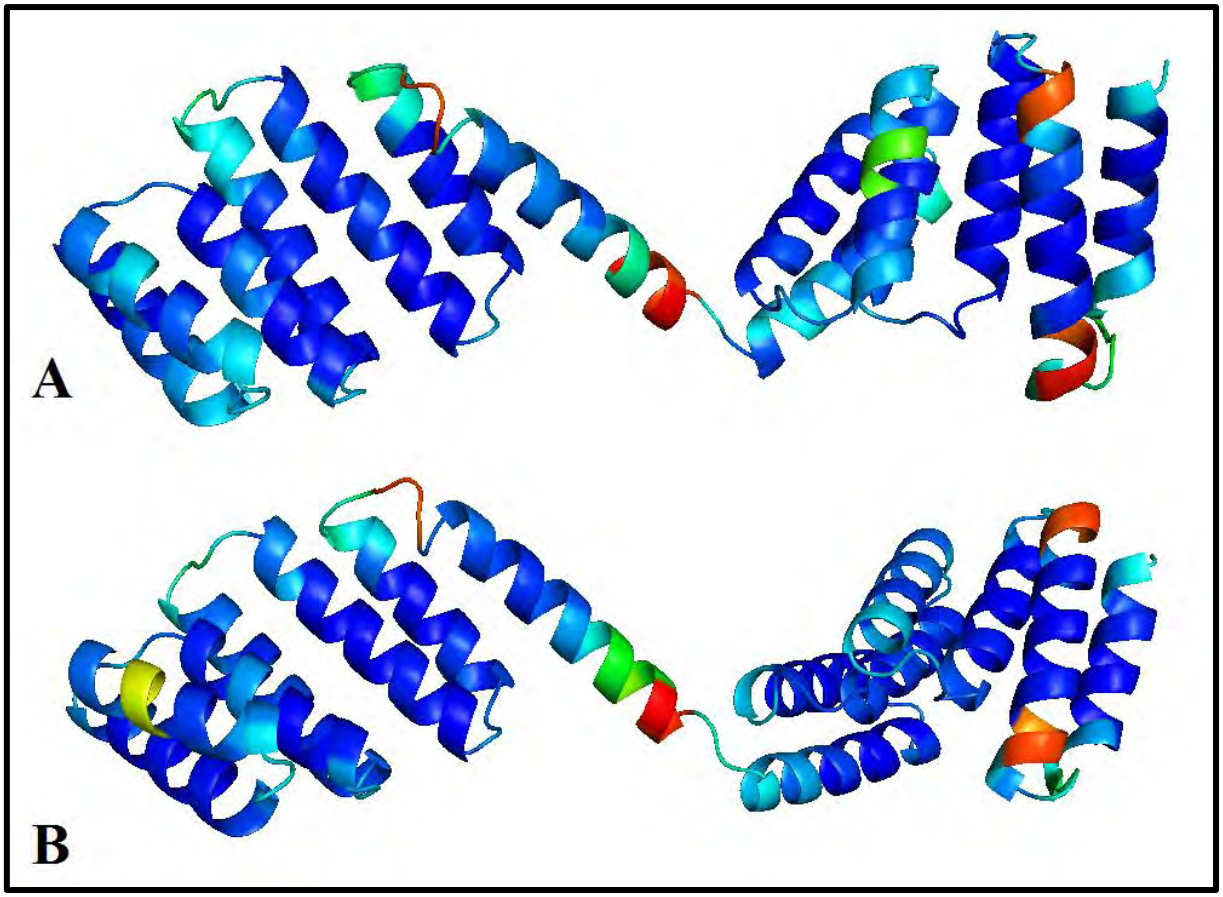


Figure 3.3: A) MetaMQAPII rendition of chain A 3UQ3 B) MetaMQAPII rendition of minimised 3UQ3 chain A.

MetaMQAPII analysis identifies several problematic regions in both the original 3UQ3 structure and the minimised version (red and orange regions, Figure 3.3). One of these regions is the tyrosine residue of the "REY clamp" within the linker (arrows, Figure 3.3 and see also Figure 3.4).

As discussed earlier in the chapter, unusual features naturally occurring within a template/ true native structure can affect the accuracy of a minimised model, as these unusual features (such as energetically unfavourable interactions) tend to be "smoothed out". There is good evidence to suggest that Hop possesses such a feature (see Chapter 1 and Chapter 2). This involves the completely conserved "REY clamp" residues which function as the rigid linker that forces TPR2AB into an S-shaped conformation (Schmid et al., 2012). In yeast, the importance of the interactions between these residues was demonstrated by mutating the arginine that positions the rigid linker (Schmid et al., 2012). This mutation affected correct folding of glucocorticoid receptor (GR) and thus GR activity was reduced to about 55%. This

rigid linker is thought to be formed through cation-π packing of the arginine side chain against the aromatic ring of the tyrosine (Figure 3.4 A). These residues are further stabilized via hydrogen bond formation between glutamine and both arginine and tyrosine. The structure of this linker was further validated through analysis of the X-ray crystallography data, revealing that electron density of the linker region between the two TPR2AB domains is structurally well defined (Schmid et al., 2012).

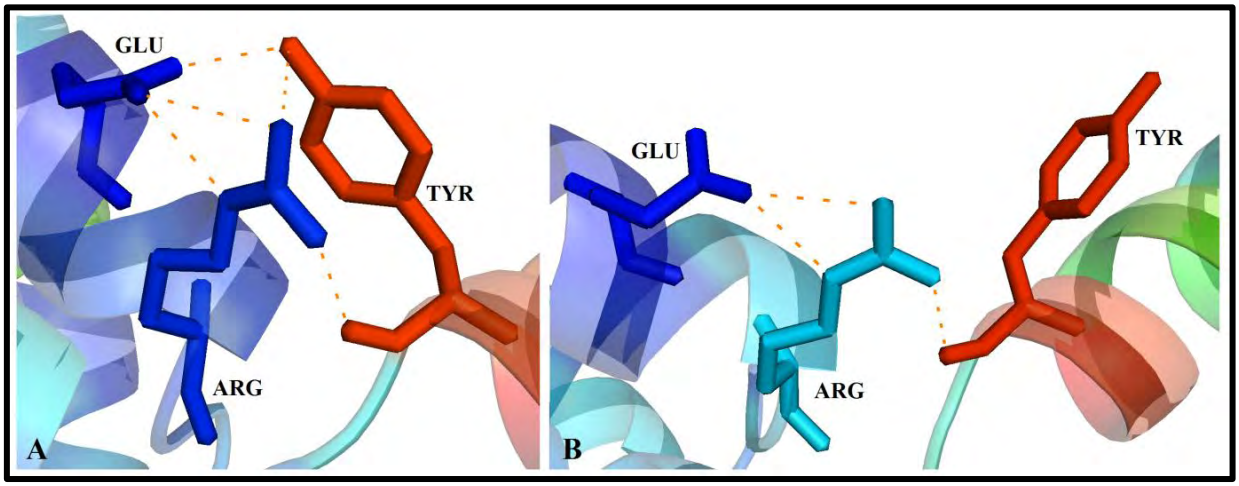


Figure 3.4: The REY clamp residues (stick representation; TYR390, GLU421 and ARG425) for MetaMQAPII rendition of chain A 3UQ3 A) Before minimisation B) After minimisation. The orange dashes indicate polar contacts predicted between the three residues in PyMOL.

From Table 3.1, it can be seen that the N-DOPE Z score and Rosetta energy for the minimised version of 3UQ3 is lower. The overall MetaMQAPII scores for both models indicate that some regions are improved while others are degraded with minimisation (see Figure 3.3). This is the case for the linker region displayed in Figure 3.4. An intra-protein interaction calculator within the PIC webserver was used to validate the existence of the REY clamp interactions in templates and models as well as their minimised counterparts. Visualisation of polar contacts within PyMOL (dotted orange lines, Figure 3.4) indicates a loss of these important interactions between these three residues within the REY clamp through reorientation of the tyrosine residue (i.e. presenting a different rotamer), while analysis of interactions predicted by the Protein Interaction Calculator (PIC, discussed in Chapter 4) see Table 3.4, indicates a total loss of interaction. These REY clamp interactions were also used as an indicator of model quality and assessed as such in all relevant templates and models.

Table 3.4: Intra-protein interactions calculated by the PIC webserver pertaining to the residues with the REY clamp. The final column represents interactions between any of the three REY residues with any non-REY residue in the structure.

	Main-Side chain Interactions	Side-Side chain Interactions	Ionic Interactions	Cation-Pi Interactions	Non intra-REY interactions
3UQ3	2 x ARG-TYR	1 x TYR-GLU 2 x ARG-TYR 3 x ARG-GLU	1 x GLU-ARG	1 x TYR-ARG	0
Minimised	0	0	0	0	0

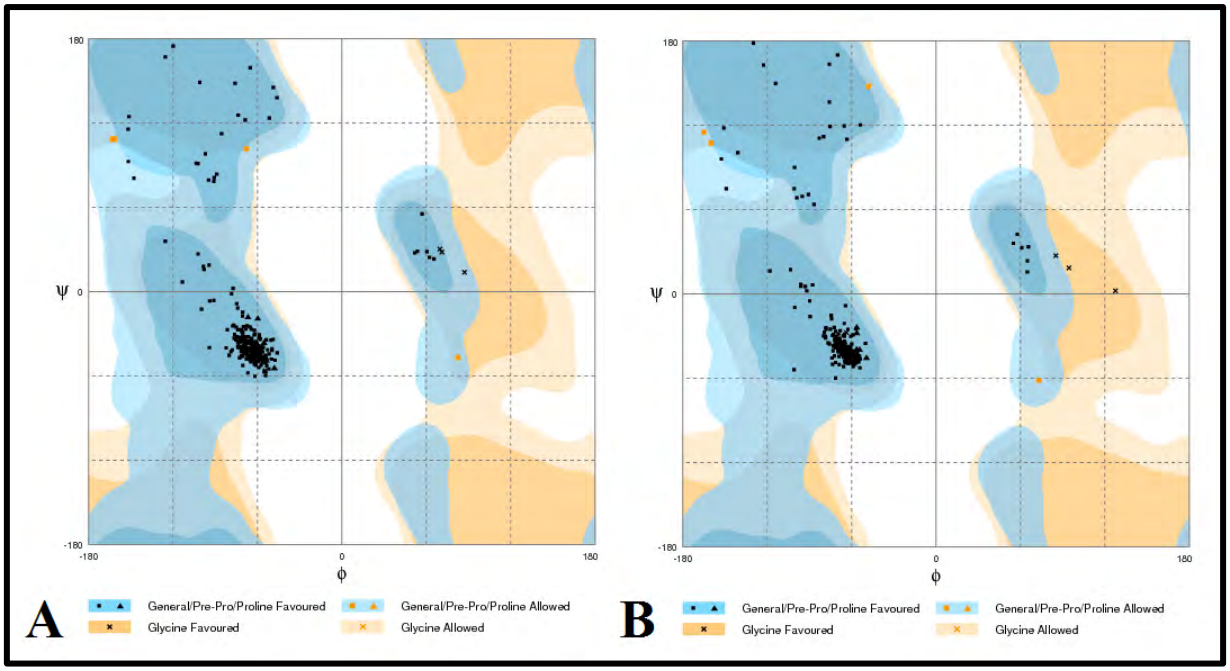


Figure 3.5: Ramachandran plots for A) 3UQ3 and B) minimised 3UQ3. Black -triangles and squares represent amino acids in the "Favoured" regions, Orange. Triangles and squares represent amino acids in the "allowed" regions.

A Ramachandran plot (Figure 3.5) indicates that both the original structure and the minimised structure are roughly of equal quality. As described above, to create models in complex with longer Hsp/Hsc70 C-terminal peptide, a second template (3UPV) was aligned to the first used to model the TPR2B region.

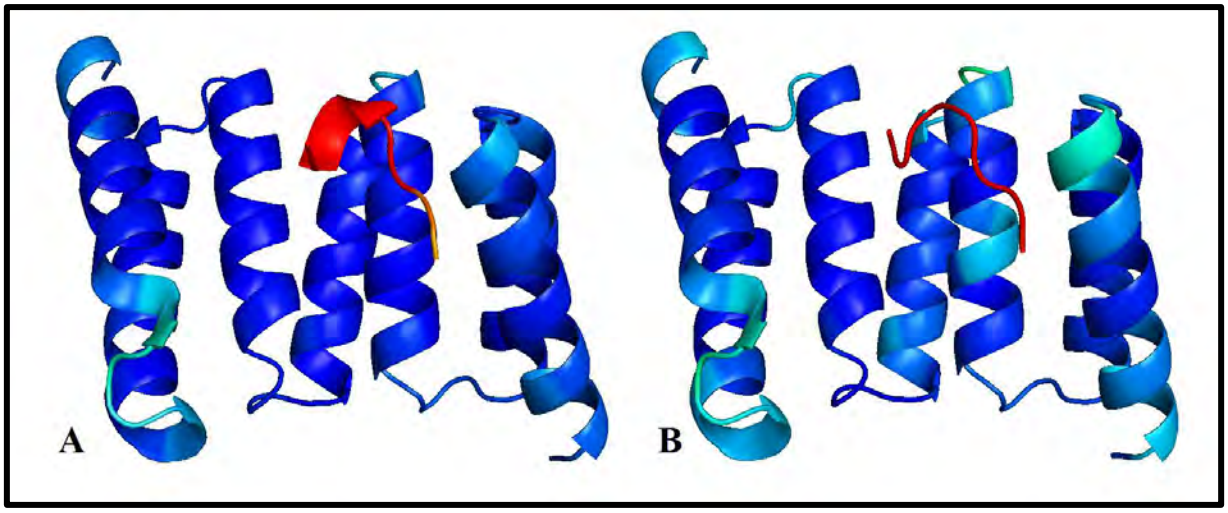


Figure 3.6: A) MetaMQAPII rendition of 3UPV. B) MetaMQAPII rendition of minimised 3UPV.

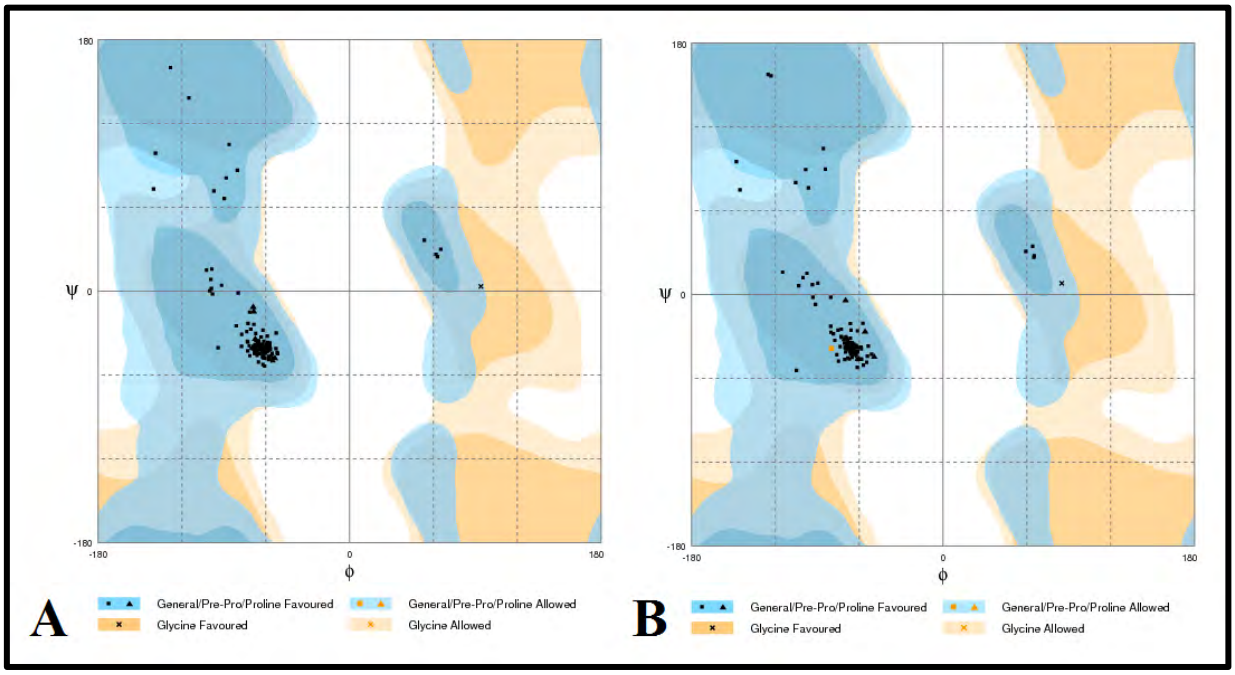


Figure 3.7: Ramachandran plots for A) 3UPV and B) minimised 3UPV. Black triangles and squares represent amino acids in the "Favoured" regions, Orange. Triangles and squares represent amino acids in the "allowed" regions.

From Table 3.1, it can be seen that the N-DOPE Z score and Rosetta energy for the minimised version of 3UPV is lower. However, the overall METAMQAPII scores for both models indicate that there is degradation of quality with minimisation, particularly for the secondary structure of the C-terminal peptide of Hsp70 (see Figure 3.6). Ramachandran plots show that the original structure is of very high quality (all residues in the "favoured" regions) and that minimisation moves a single residue (PHE312) from the "favoured" to the "allowed" region (see Figure 3.7).

3.3.2.3 Templates for Modelling the ScHsp90 M and C Domains in Complex with HopTPR2

Schmid et al. (2012) and (Lee et al., 2012) have suggested that as the affinity of Hop for Hsp90 is much higher than that of the Hsp90 C-terminal peptide. There is most likely additional binding site/s for Hop in Hsp90, especially considering that the ATPase inhibition by Hop also requires an additional interaction site/s (Lee et al., 2012). They thus investigated this interaction by docking the ScHop TPR2 fragment (3UQ3, minus C-terminal partner peptides) to a low-resolution spin-labled model ScHsp90 M and C domain fragment. Spin labelling was done with three single cysteine variants of the ScHsp90-M domain. As ScHsp90 does not possess cysteine, three serines (at postions 411, SER422, SER456) were mutated to cysteine which were then modified via the addition of iodoacetamido-proxyl and reduced with ascorbic acid. The proxyl-modified template of the Hsp90 M and C domain fragment was refined with software that back calculates NOE restraints from a crystal structure template (Schwieters et al, 2003). The template used to do this was 2CGE, which is a ScHp90 M and C domain fragment in complex an ATP analogue and the co-chaperone Sba1 (Ali et al., 2006).

As such, this complex was not published in the PDB, however the coordinates for this structure were kindly provided by the authors (see file 'hsp90_sti1_clomplex.pdb', the modified cysteine residues are identified as CYM). As a template, the structure presented several issues. This structure does not address conformational changes in the overall backbone structure of the Hop TPR2 fragment owing to C-terminal peptide binding, versus that resulting from alternate Hsp90 binding. The modified CYM residues in the Hsp90 half of the complex had to be changed to either cysteine (mutant) or serine (native), owing to constraints in the Modeller software, which does not recognise non-standard amino acids. Two of these modified CYM residues are in the region where Hsp90 interacts with HopTPR2 (see Figure 3.8).

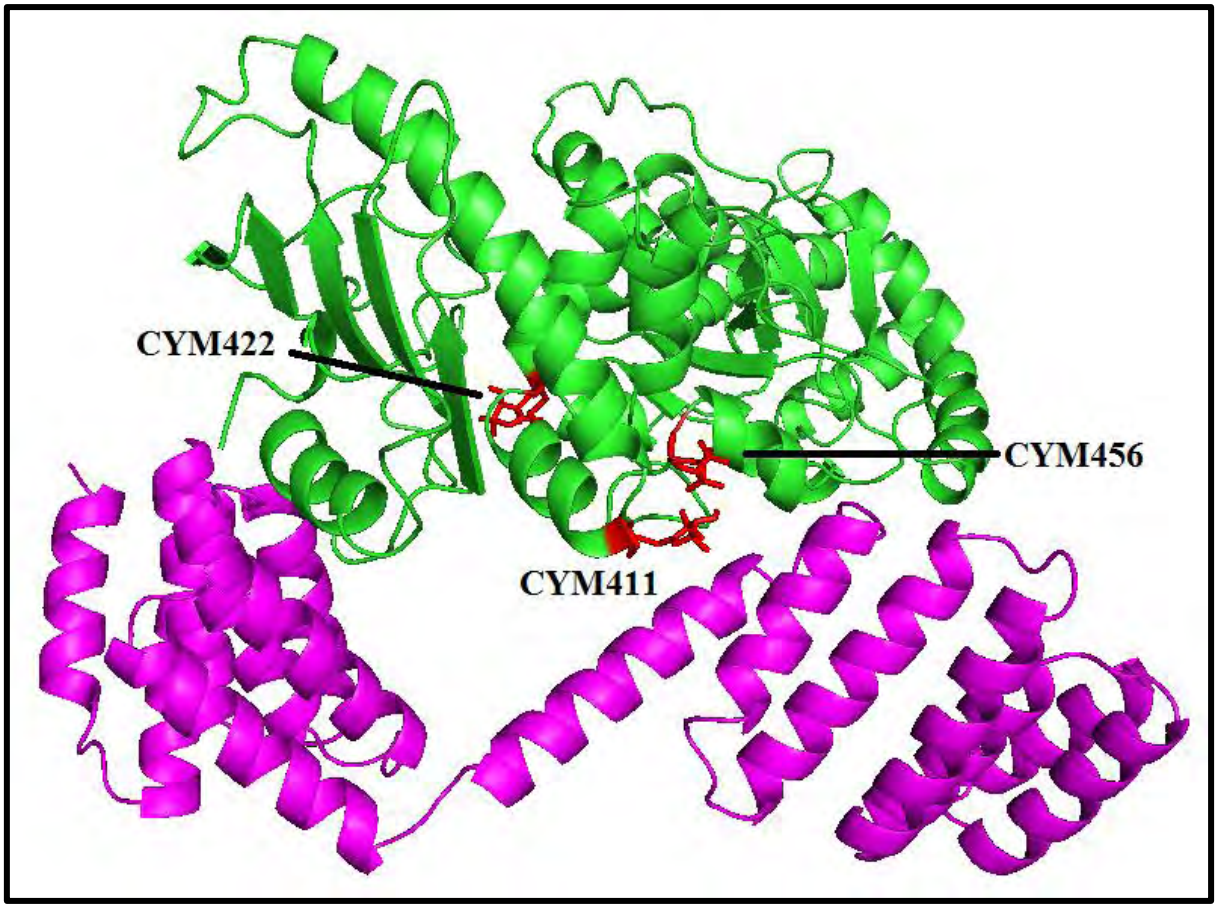


Figure 3.8: Cartoon representation of Hop TPR2 (magenta) in complex with Hsp90 M and C domain (green). The three proxyl modified cysteines are labelled and displayed in stick representation (scarlet). CYM411 and 456 are close to the site of interaction with TPR2A.

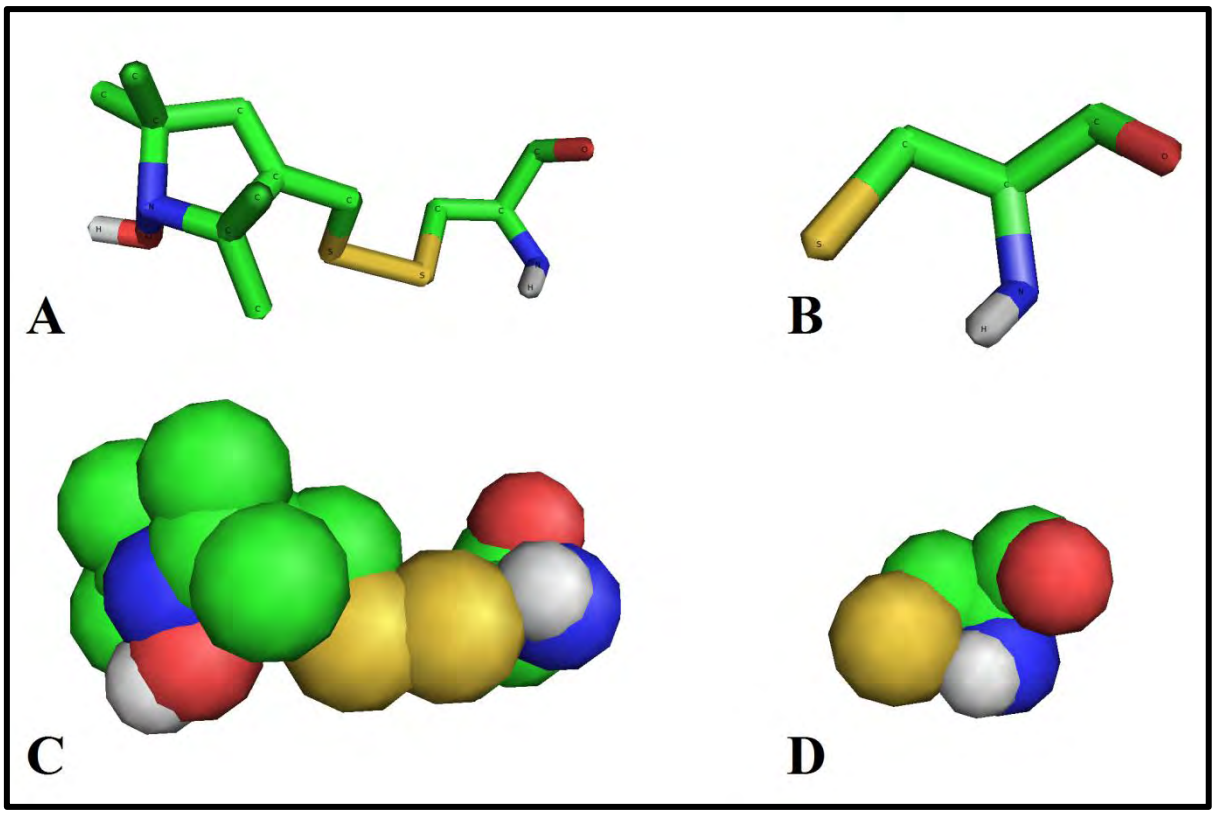


Figure 3.9: Stick representation of A) the proxyl modified cysteine (CYM) and B) cysteine displayed above the sphere representation of C) CYM and D) cysteine. The proxyl group is a cyclic, bulky molecule that may lead to distortion of the native protein backbone.

It was decided that *in silico* modification of the model would be limited to removing the atomic co-ordinates of the proxyl group from the modified cysteines, leaving the template model as a cysteine mutant (see Figure 3.9 and file 'tailess_schimdCYS.pdb'). This is because there is less modification of the template than that required to modify to serine (which could introduce significant backbone errors). PyMOL predicted no protein-protein polar contacts for the three mutant CYS residues in the modified template file, while the PIC predicted a single side-side chain hydrogen bond interaction between the sulfur atom of CYS411 (H-bond donor) and the second side-chain oxygen of GLU379 in HopTPR2. Both of these atoms were detected by alanine scanning and were predicted to contribute favourably to complex formation.

As part of the homology modelling control process, best of 100 self-models were analysed for each template. For this template, two self-models were built on the template structure 'tailess_schmidCYS'; the sequences for these differed only by the three residues representing the native structure (with serine in place of CYM) and the mutant structure (with cysteine in place of CYM). It is interesting to note that the best of 100 self-models representing the native

Hsp90 had lower N-DOPE Z, Rosetta energy and Cα-RMSD (to the template) scores, than the best of 100 representing mutant Hsp90 (see Table 3.5). This demonstrates that just three cysteine mutations can noticeably affect overall backbone structure compared to the native structure of Hsp90.

Table 3.5: Comparing native versus mutant self-models of the template 'tailess_SchmidCYS'.

Model	N-DOPE Z	Rosetta	Cα-RMSD
tailess_schmidCYS.pdb	-1.069	199.047	0.000
tailess_schmidCYS_cysteinecontrol.B99990098.pdb	-0.879	1829.022	0.242
tailess_schmidCYS_serinecontrol.B99990024.pdb	-0.905	1777.032	0.209

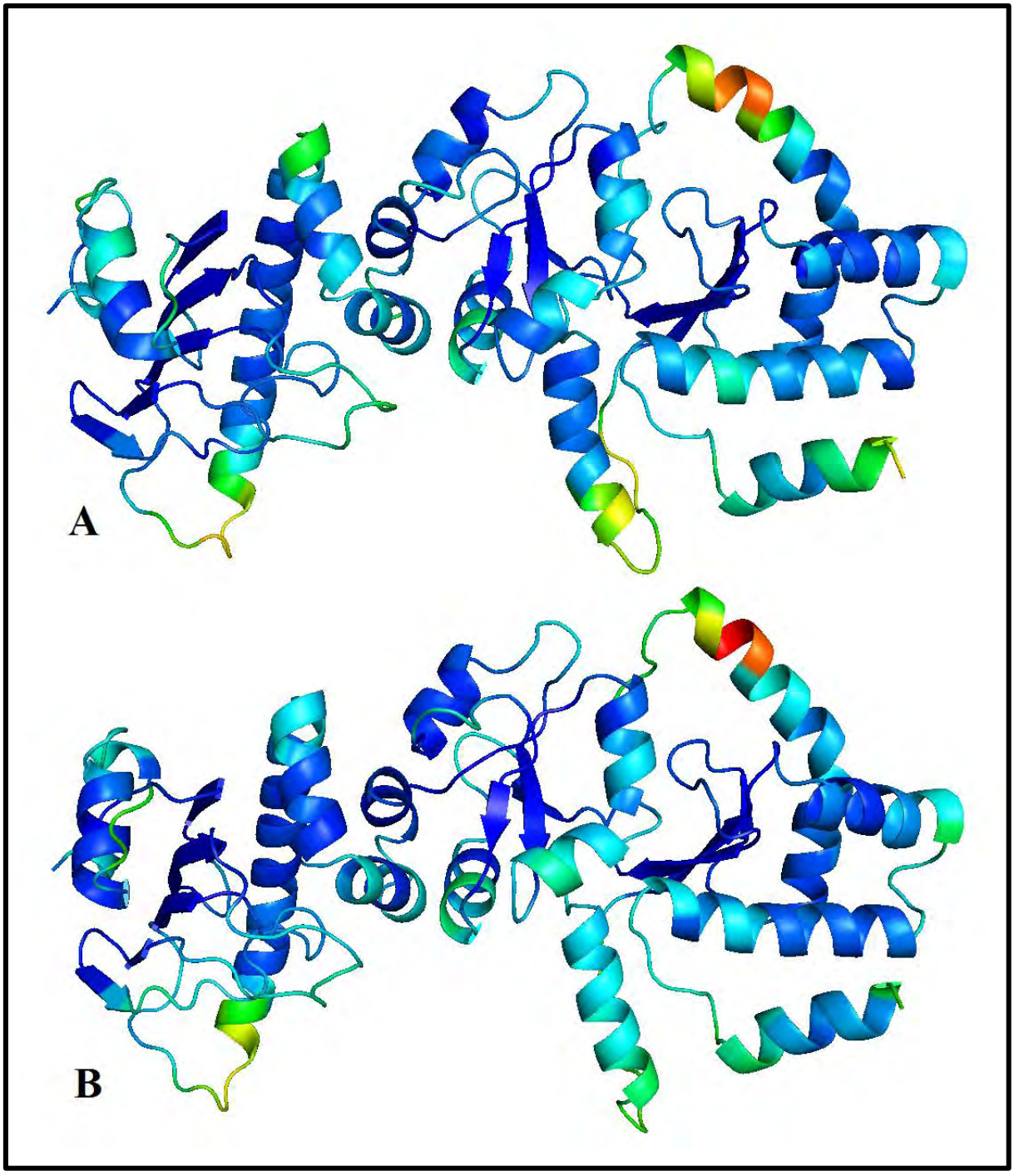


Figure 3.10: A) MetaMQAPII rendition of the Hsp90 half of SchmidCYS. B) MetaMQAPII rendition of the minimised Hsp90 half of SchmidCYS.

For general analysis of the 'tailess_schmidCYS' template, it was plain to see from Table 3.1 that minimisation resulted in marked improvement of the N-DOPE Z and Rosetta energy scores. Based on MetaMQAPII scores, the Hsp90 half of the complex shows overall improvement with minimisation (see Figure 3.10), which is confirmed by the Ramachandran plots for the original and minimised versions of Hsp90 (see Figure 3.11). Minimisation reduced the number of residues occupying both the disallowed and allowed phi and psi

regions, and increased the percentage of residues within the favoured region from 91.4% to 93.4%.

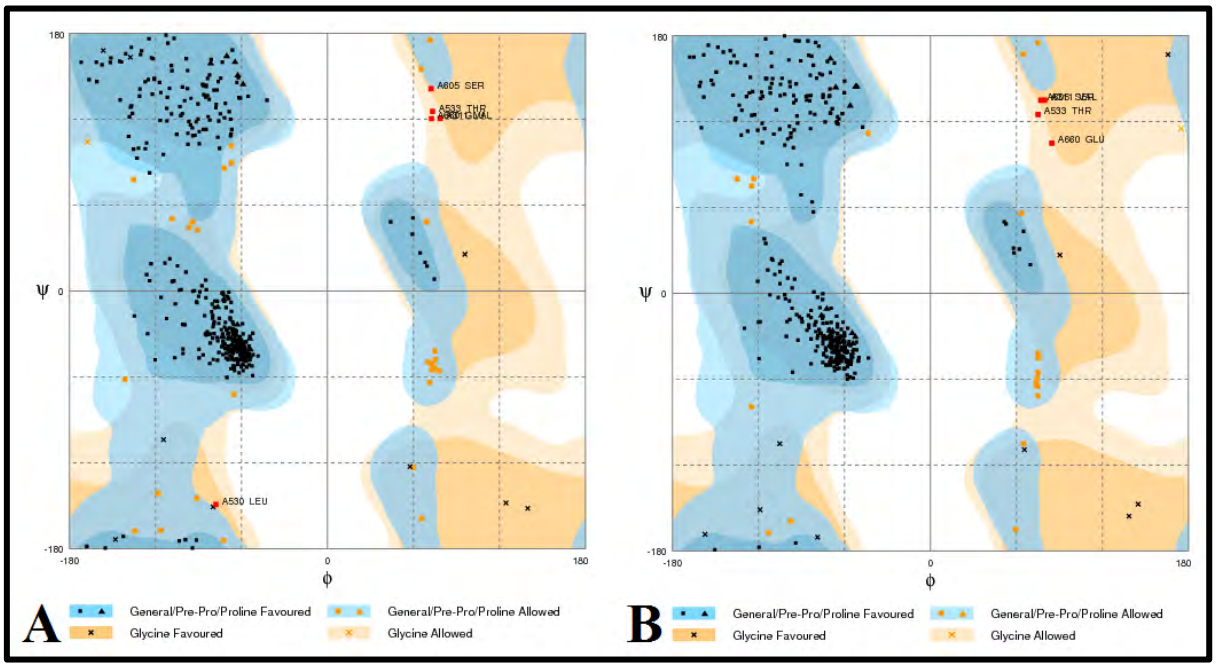


Figure 3.11: Ramachandran plots for A) Hsp90 half of SchmidCYS, red squares indicate residues (VAL311, LEU530, THR533, SER605 and GLU660) occupying disallowed regions. B) Minimised Hsp90 half of SchmidCYS, red squares indicate residues (VAL311, THR533, SER605 and GLU660) occupying disallowed regions. Black triangles and squares represent amino acids in the "Favoured" regions, Orange. Triangles and squares represent amino acids in the "allowed" regions.

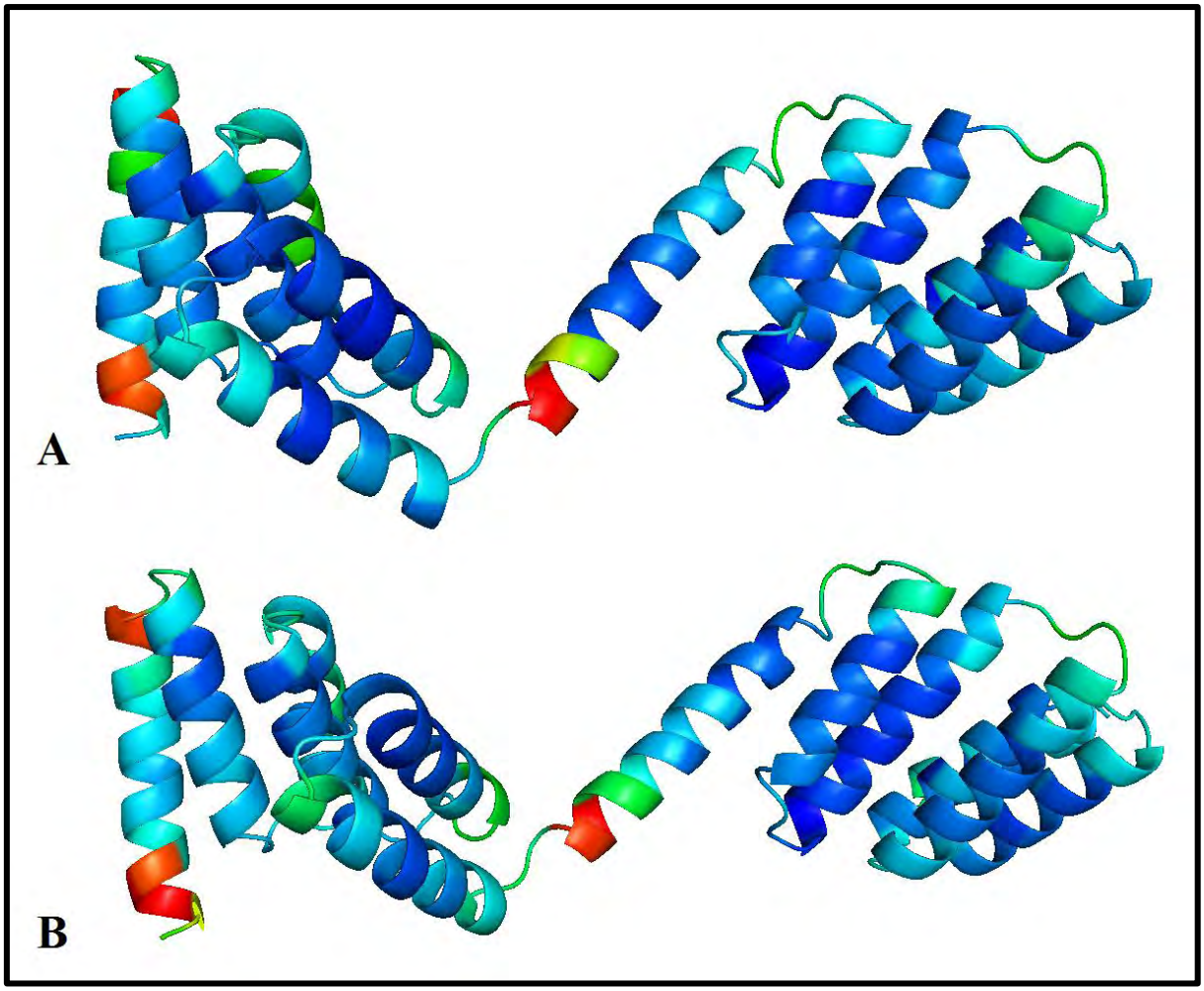


Figure 3.12: MetaMQAPII rendition of the Hop half of SchmidCYS. B) MetaMQAPII rendition of the minimised Hop half of SchmidCYS.

Based on MetaMQAPII representation, the Hop half of the complex appears to decrease in quality with minimisation (see Figure 3.12), while Ramachandran plots indicate that both the original and minimised versions of Hop are of equal quality (see Figure 3.13).

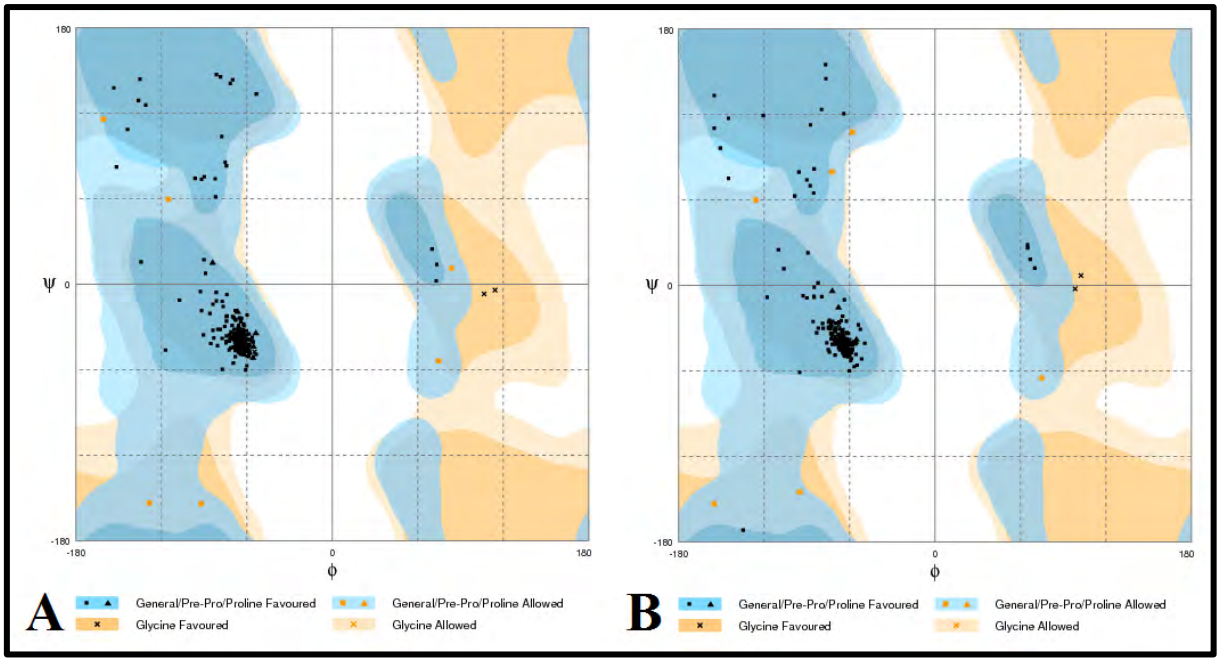


Figure 3.13: Ramachandran plots for A) Hop half of SchmidCYS and B) minimised Hop half of SchmidCYS. Black triangles and squares represent amino acids in the "Favoured" regions, Orange. Triangles and squares represent amino acids in the "allowed" regions.

Table 3.6: Intra-protein interactions calculated by the PIC webserver pertaining to the residues with the REY clamp in the Hop half of SchmidCYS. The final column represents interactions between any of the three REY residues with any non-REY residue in the structure.

	Main-Side chain Interactions	Side-Side chain Interactions	Ionic Interactions	Cation-Pi Interactions	Non intra-REY interactions
Hop half of tailess_ schmidCYS	2 x ARG-TYR	2 x TYR-GLU 2 x ARG-TYR 3 x ARG-GLU	1 x GLU-ARG	1 x TYR-ARG	1 x Ionic GLU-LYS424 1 x Cation-pi TYR-LYS424
Minimised	0	0	0	0	0

MetaMQAPII and PIC analysis of the REY clampin the Hop half of SchmidCYS, shows a complete degradation of interactions between the three residues (see Figure 3.14 and Table 3.6) as well as a change in orientation of the tyrosine with respect to glutamine and arginine. This is similar to change in the REY clamp displayed in the minimised version of 3UQ3. It is

interesting to notice that Table 3.6 and 3.4 show slightly different results, considering that the Hop half of SchmidCYS is 3UQ3. This difference in co-ordination of the residues in the REY clamp was most probably a result of fine backbone changes that occurred due to docking algorithm (and likely several refinement steps) used to create the model in the first place. Alternatively, this difference may also have been as a result of fine backbone changes to the S-shaped conformation of TPR2A when not bound to both Hsp90 and Hsp70 C-terminal peptides or when bound to Hsp90 M and C domain. As there was slightly more interactions between the REY clamp residues in Table 3.6, this may indicate that alternate binding to Hsp90 M and C domains actually stabilises or promotes the S-shaped conformation of the TPR2 domain when in the multi-chaperone complex, a conclusion arrived at elsewhere in the literature (Cheung-Flynn et al., 2003; Schmid et al., 2012; Southworth & Agard, 2011).

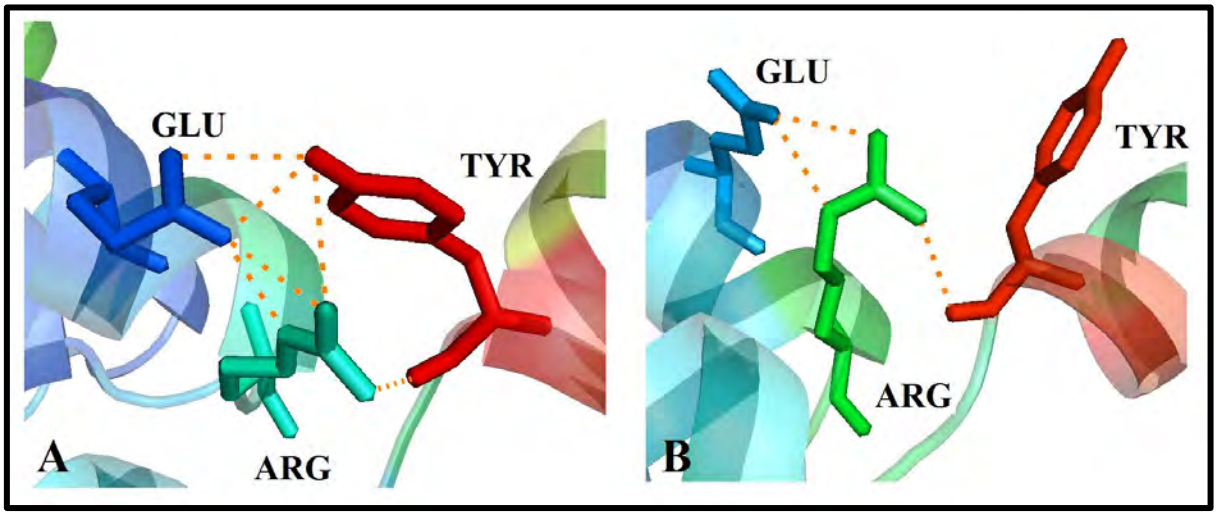


Figure 3.14: The REY clamp residues (stick representation; TYR390, GLU421 and ARG425) for MetaMQAPII rendition of chain B SchmidCYS. A) Before minimisation and B) after minimisation. The orange dashes indicate polar contacts predicted between the three residues in PyMOL.

Recently, two NMR structures of *C. elegans* N-terminal Hop TPR regions in isolation were published to the PDB (Osipiuk et al., 2012). Owing to initial difficulties in modelling PfHop TPR2 in complex with PfHsp90 M and C domains, several multi template combinations were explored. One of these structures from CeHop (4GCO, with high similarity to the TPR2B region in PfHop), in combination with the "tailess_schmidCYS"template, yielded the best homology models. Based on the analysis of several pilot models of PfHop in complex with PfHsp90, the TPR2B region adopts a wider range of conformations than does the TPR2A region. The slightly better model produced with the inclusion of 4GCO as a template may

account for small conformational changes that better reflect the PfTPR2B in complex with Hsp90 while not bound to C-terminal peptide of Hsp70. This is because in this structure, 3UQ3 is not represented in complex with the C-terminal peptides of Hsp70 and Hsp90, with which its structure was originally determined, while the structure for 4GCO was determined unbound to C-terminal peptide (Osipiuk et al., 2012). This is speculative and it is possible that the model quality was increased simply because 4GCO is a better quality template (see Table 3.2) than 3UQ3 (see Table 3.1). However, it is curious to note that 3UPV, which has similar template quality to 4GCO (see Table 3.1), did not yield models of comparable quality in place of 4GCO. It is more likely as a result of subtle sequential or structural features that are shared by invertebrates and protozoans but not by fungi, such as motif 9 (olive green, Figure 2.6), discussed in Chapter 2.

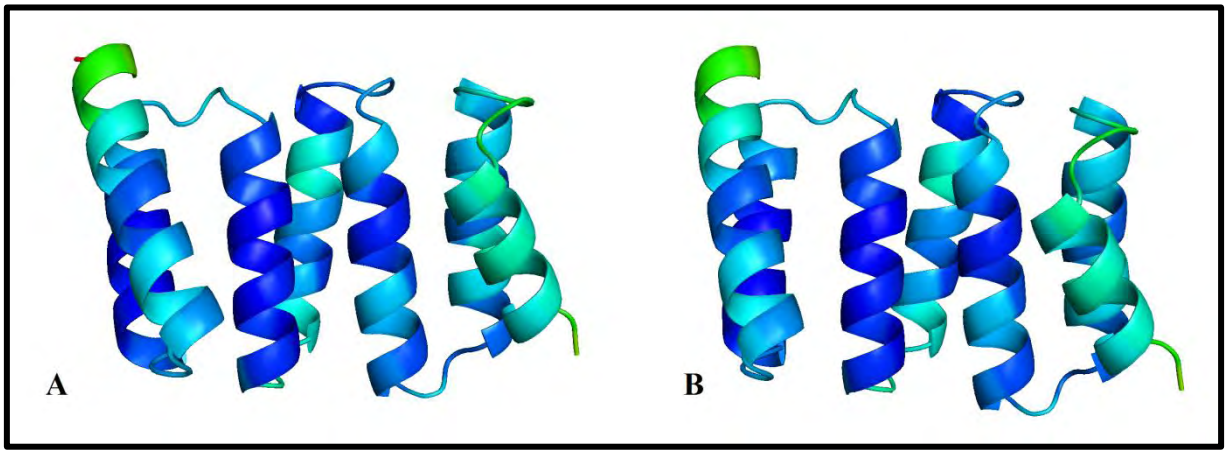


Figure 3.15: A) MetaMQAPII rendition of 4GCO. B) MetaMQAPII rendition of the minimised 4GCO template.

Overall analysis, shows that 4GCO was an excellent template structure that minimisation did little to improve on (see respective N-DOPE Z and Rosetta energy scores, Table 3.2). MetaMQAPII analysis shows that there was almost no change of 4GCO with minimisation (Figure 3.15); while Ramachandran plot analysis indicated that the minimised and original versions of the template were of equal quality (Figure 3.16).

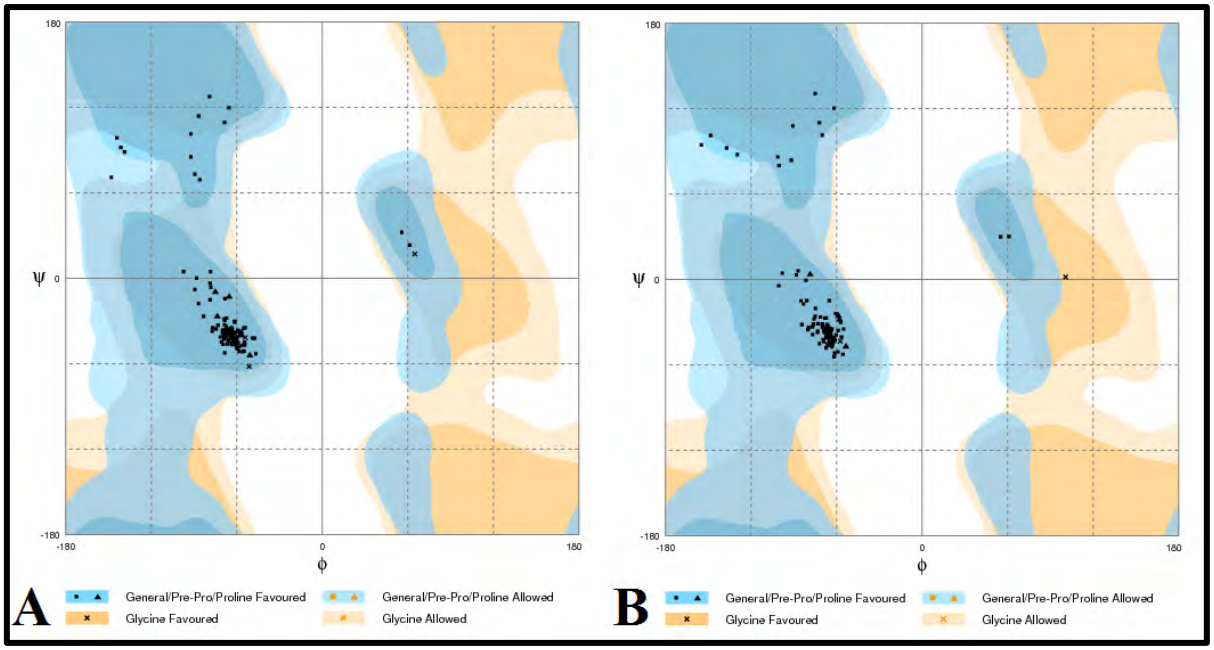


Figure 3.16: Ramachandran plots for A) 4GCO and B) minimised 4GCO. Black triangles and squares represent amino acids in the "Favoured" regions, Orange. Triangles and squares represent amino acids in the "allowed" regions.

3.3.2.4 Template Analysis for DP Structures

As seen from Table 3.1, the average of 21 NMR models used to model DP1 in HsHop and PfHop was of high quality, with low N-DOPE Z and Rosetta energy scores, both before and after minimisation. This was because DP1 appeared to have a very rigid structure and all 21 models were in good agreement, so the average was well represented with the data (see Figure 1.5, Chapter 1). It was for this reason that the DP1 regions in both PfHop and HsHop were modelled based on a minimised average of the 21 models in 2LLV.

For DP2, however, there was slightly weaker agreement of the models (see Figure 1.5, Chapter 1). This disagreement was thought to explain the functionality of DP2 as a cleft that helps stabilize client protein, and the models seem to represent the range from most open to most closed conformation of DP2 (Schmid et al., 2012). A different approach was used to select a working template, as the average structure of all 21 models had very high N-DOPE Z and Rosetta energy scores, and even minimisation of the average yielded a poor model (Table 3.2). Based on the assumption that the minimised average is a good proxy for the native, unbound, solution state of DP2 (the rationale for this is explained in detail in Chapter 4), a single model out of the 21 was selected based on how well it was structurally aligned to the

minimised average. The model selected was that representing state 9 in PDB entry 2LLW (see Figure 3.17).

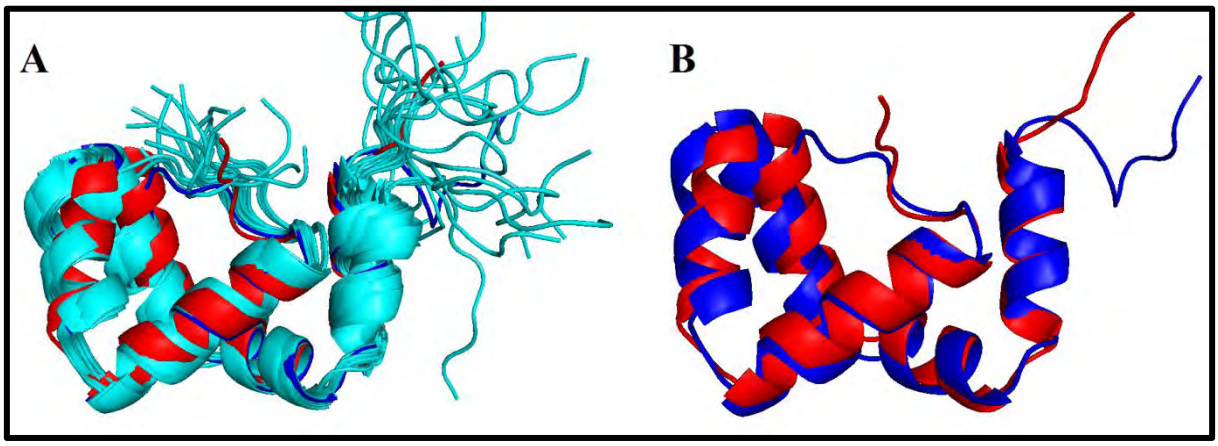


Figure 3.17: Selecting a representative template for DP2. A) All 21 models in 2LLW are superposed and represented in cartoon (cyan), with the exception of B) the minimised average of all models (scarlet) and state 9 (navy).

It was satisfying to note that the selected state had the lowest N-DOPE Z score (-2.166) of all 21 models. It also appears that the minimised average makes a good proxy for the native and unbound solution state of DP2, as it represented the most closed (globular) conformation state of 2LLW. As discussed in Chapter 1, this closed conformation is most likely because of shielding of the hydrophobic residues in the centre of the protein (see Figure 1.8.D, Chapter 1) that may assist in interaction with Hsp70's client protein.

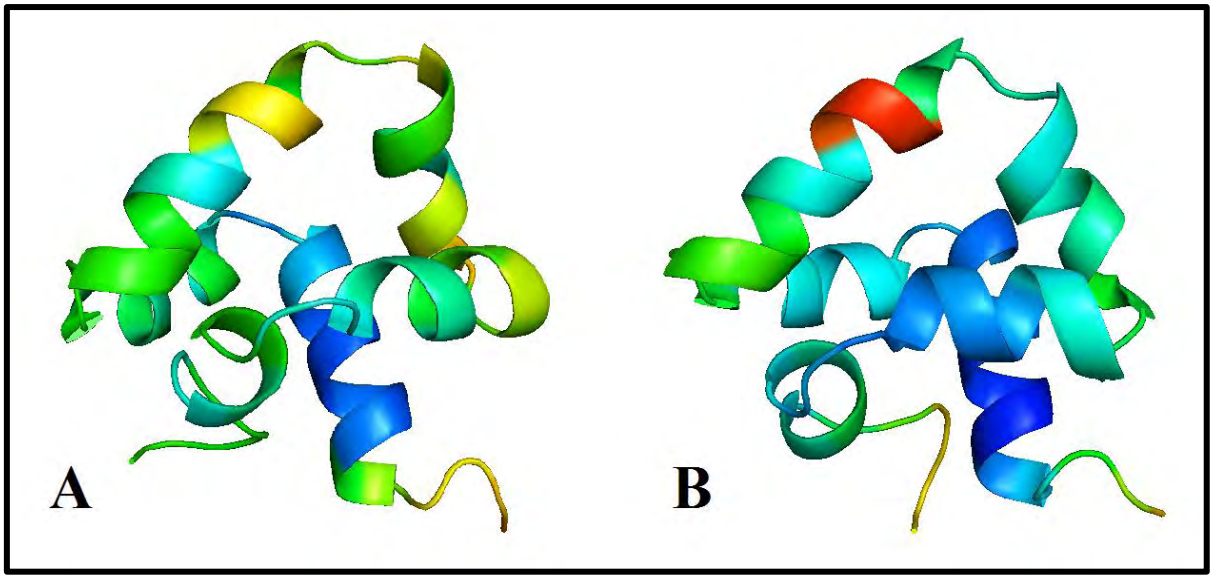


Figure 3.18: A) MetaMQAPII rendition of the average structure of 2LLV B) MetaMQAPII rendition of the minimised average structure of 2LLV.

MetaMQAPII analysis of both the average and minimised average structure of both 2LLV (Figure 3.19) and 2LLW (Figure 3.18 A and B) indicated that the structures are slightly degraded with refinement in certain regions while other regions are improved. The MetaMQAPII scores for state 9 indicated poorer model quality (Figure 3.19), which was surprising considering that the N-DOPE Z and Rosetta energy scores (-2.166 and 83.419, respectively) for state 9 are drastically lower than that of the average and minimised average structures (see Table 3.1).

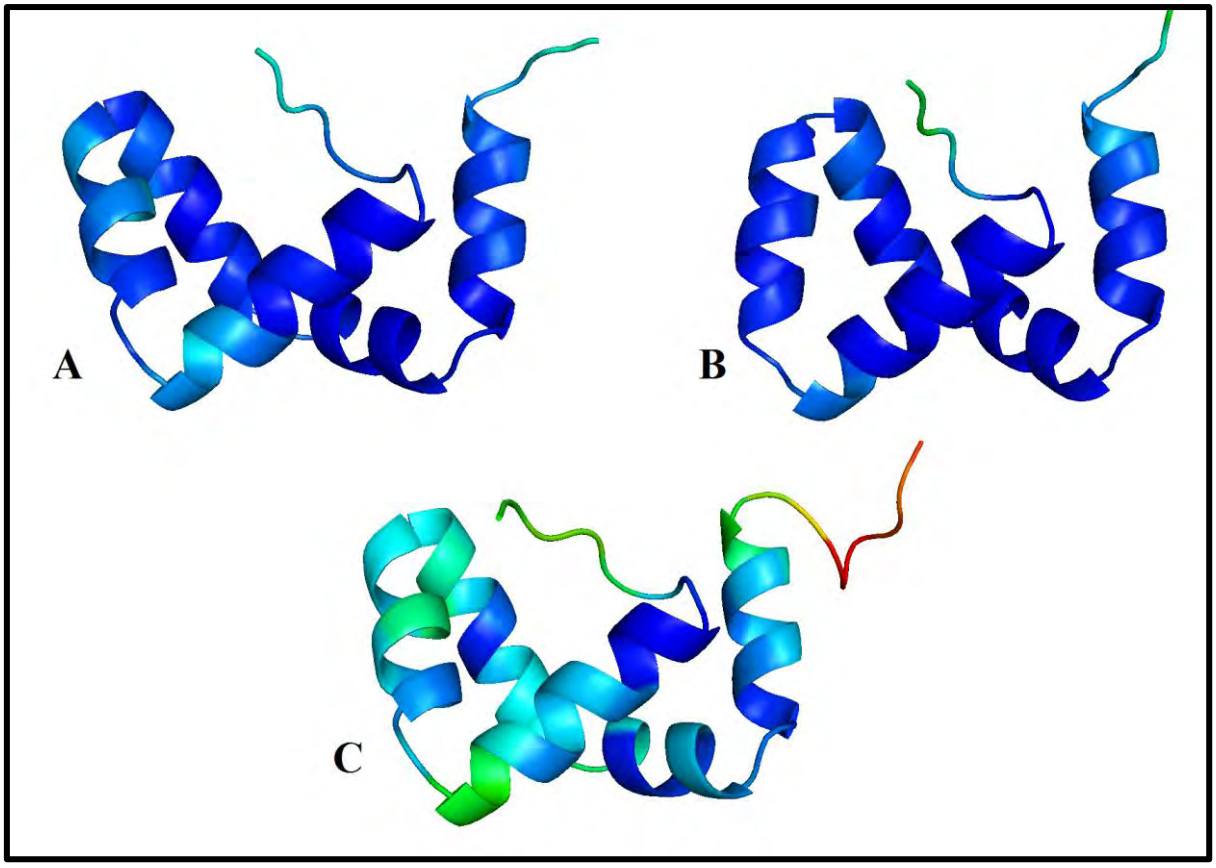


Figure 3.19: A) MetaMQAPII rendition of the average structure of 2LLW. B) MetaMQAPII rendition of the minimised average structure of 2LLW. C) MetaMQAPII rendition of the state 9 of 2LLW.

3.3.3 Homology Modelling Summary

Table 3.7 below summarises the energy scores for the best of 100 self-models built for each template. This was a form of control to test the efficacy of the homology modelling process. For almost all self-models built, the N-DOPE Z-scores are very close to that of the original templates (see Tables 3.1 and 3.2) and are in general very good models (N-DOPE Z > 1).The exceptions are the two SchmidCYS models, which have significantly lower N-DOPE Z scores than the original template (the reasons for this are discussed in Section 3.4.3) and the 3UQ3 self-model, which has a relatively large Cα-RMSD score compared to its original template. This is likely as a result of problems in modelling the constrained linker region (discussed in Section 3.3.2.2 See also Section 2.6.8, in Chapter 2) that is responsible for the S-shaped backbone conformation in the structure of TPR2.

Table 3.7: Energy scores for best of 100 self-models built for each template.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
4GCO_mod_control.B99990013.pdb	-1.973	46.339	0.146
tailess_schmidCYS_cysteinecontrol.B99990098.pdb	-0.879	1829.022	0.242
tailess_schmidCYS_serinecontrol.B99990024.pdb	-0.905	1777.032	0.209
1ELW_mod_control.B99990017.pdb	-2.3243	-105.552	0.126
2LLV_average_control.B99990072.pdb	-1.677	471.713	0.269
2LLW1_state009_control.B99990020.pdb	-2.147	250.841	0.320
3uq3_TPR2ab_control.B99990073.pdb	-1.304	5.464	0.873

Table 3.8 summarises the energy scores for the best of 100 models built for each species, and, if the case, the minimised version of that model. These are ultimately the final products of Chapter 3. The files for these structures have been provided in the supplementary data provided on disk and their validation has been described in Sections 3.3.3.1 and 3.3.3.2.

Table 3.8: Energy scores for best of 100 models built for each species domain or interaction.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
Minimised_PfHopTPR1.B99990036_1.pdb	-2.432	-361.134	6.926
PfHopTPR1N.B99990050.pdb*	-2.135	261.725	6.768
HsHopTPR1N.B99990001.pdb*	-2.342	-69.055	0.103
HsTPR2ab2yeast.B99990013.pdb	-0.919	980.277	10.984
HsTPR2ab2N.B99990083.pdb*	-0.976	774.873	10.938
PfTPR2ab2yeast.B99990014.pdb	-1.056	1248.952	7.180
PfTPR2ab2N.B99990061.pdb*	-1.072	1109.672	7.028
MinimisedHscomplex_07_1.pdb	-1.095	-1600.531	5.850
MinimisedPfMulti_102.pdb	-0.824	-1625.737	15.991
PfDP1.B99990060.pdb	-0.786	604.815	11.490
PfDP2009.B99990003.pdb	-2.029	446.246	0.490
HsDP1.B99990080.pdb	-1.623	485.168	9.265
HsDP2.B99990038.pdb	-2.343	62.263	5.879

* Alternate PfHsp70-x C-terminal binding complexes.

It is important to note that the four models with an "N" in the file name, are the models created to simulate the interactions between TPR regions in HsHop interacting with the alternate C-terminal motif present in the Hsp70 variant, PfHsp70-x, that is transported to the erythrocyte cytosol (as discussed in Chapter 1, Section 1.7 and Chapter 4, Section 4.5). These validation for these models is not discussed owing to time constraints, however their quality is very similar to that of their normal counterparts (i.e. TPR domains in complex with Hsp70-1), as the models differ by one or two residues.

3.3.3.1 Model Validation for Complexes Involving TPR Motifs

Just as for the templates, the best models were evaluated in terms of N-DOPE Z scores, Rosetta energies, Cα-RMSD scores, MetaMQAPII scores, Ramachandran plots and important structural features. A good overview of the quality of homology models was critical for understanding the limitations of these models, in order to correctly interpret results in later protein-protein interaction studies.

3.7.1 Analysing models involved in HopTPR1:Hsp70 interactions

Table 3.9: Energy scores for top 10 PfHopTPR1:Hsp70-GPTVEEVD complex models. Models are listed in order of decreasing N-DOPE Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
1ELW_mod.pdb	-2.569	-257.680	0.000
Minimised_PfHopTPR1.B99990036_1.pdb	-2.432	-361.134	6.926
PfHopTPR1.B99990036.pdb	-2.117	247.527	6.876
PfHopTPR1.B99990029.pdb	-2.106	203.831	6.834
PfHopTPR1.B99990060.pdb	-2.091	280.568	6.770
PfHopTPR1.B99990030.pdb	-2.071	245.246	6.871
PfHopTPR1.B99990021.pdb	-2.067	194.782	6.911
PfHopTPR1.B99990027.pdb	-2.061	277.793	6.817
PfHopTPR1.B99990095.pdb	-2.060	422.257	6.929
PfHopTPR1.B99990087.pdb	-2.057	286.809	6.871
PfHopTPR1.B99990005.pdb	-2.047	292.081	6.751
PfHopTPR1.B99990055.pdb	-2.039	306.761	6.807

From Table 3.9, it is easy to see that the overall model quality for the top 10 models is high (N-DOPE Z<<< -0.5).The Cα-RMSD score is slightly increased after minimisation, reducing the backbone similarity of the model to the template. This is a small trade-off for a better quality model and hence, both the original and the refined models were used for further analysis.

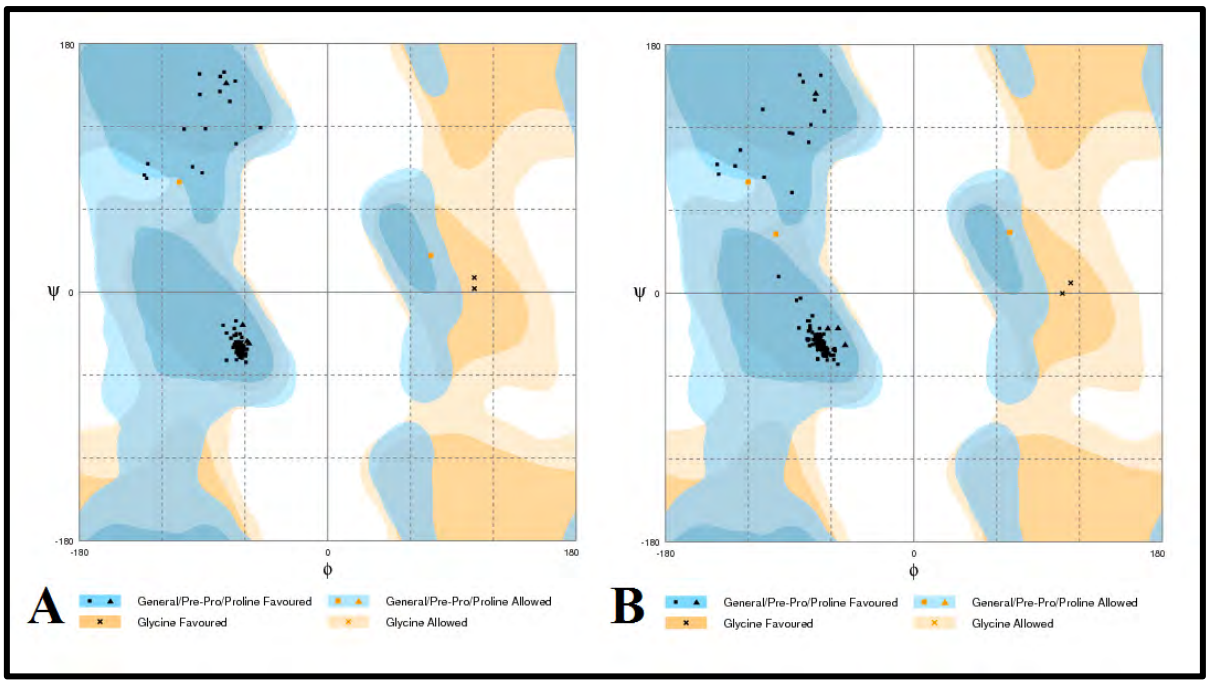


Figure 3.20: Ramachandran plots for PfTPR1 model 36. A) Original model plot. B) Minimised model plot. Black triangles and squares represent amino acids in the "Favoured" regions, Orange. Triangles and squares represent amino acids in the "allowed" regions.

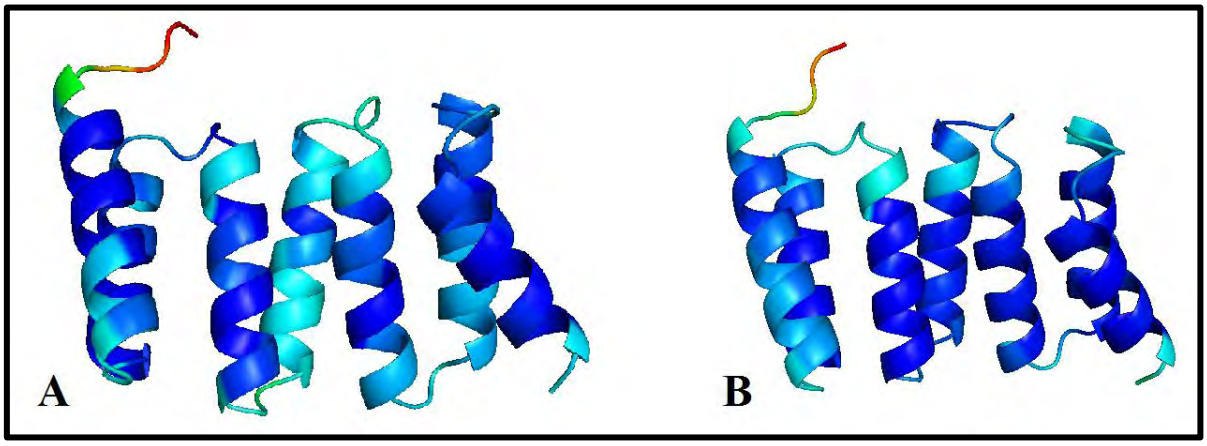


Figure 3.21: A) MetaMQAPII rendition of chain A for PfTPR1 model 36. B) MetaMQAPII rendition of minimised PfTPR1 model 36.

Refinement of the PfTPR1 model 36 returns the lower N-DOPE Z and Rosetta energy score, at the cost of moving a single residue (ASP73) from the "favoured" to the "allowed" region in the Ramachandran plot (Figure 3.20). However, MetaMQAPII analysis indicates overall improvement with minimisation (Figure 3.21).

3.3.3.2 Analysing Models Involved in Hsp90:HopTPR2A&B:Hsp70 Interactions

In the next few pages, the assessment results of the best of the Hsp90:HopTPR2A&B:Hsp70 complex models in human Hop are discussed.

Table 3.10: Energy scores for the top 10 HsHsp90-MEEVD:HsHopTPR2AB:HsHsp70-PTIEEVD complex models. Models are listed in order of decreasing DOPE-Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
Template 1: 3UPV_modH.pdb	-2.376	-159.663	9.272
Template 2: 3uq3_TPR2ab.pdb	-1.562	-229.453	0.000
Minimised_HsTPR2ab2yeast.B99990013_1.pdb	-1.253	-728.822	10.847
HsTPR2ab2yeast.B99990013.pdb	-0.919	980.277	10.984
HsTPR2ab2yeast.B99990038.pdb	-0.915	1104.424	10.863
HsTPR2ab2yeast.B99990046.pdb	-0.886	1168.660	11.027
HsTPR2ab2yeast.B99990027.pdb	-0.855	1314.817	11.846
HsTPR2ab2yeast.B99990010.pdb	-0.840	1450.146	11.488
HsTPR2ab2yeast.B99990067.pdb	-0.834	1023.708	10.816
HsTPR2ab2yeast.B99990080.pdb	-0.831	1103.644	11.479
HsTPR2ab2yeast.B99990048.pdb	-0.828	1037.730	11.247
HsTPR2ab2yeast.B99990034.pdb	-0.826	1430.663	11.347
HsTPR2ab2yeast.B99990091.pdb	-0.824	1121.761	10.872
HsTPR2ab2yeast.B99990082.pdb	-0.821	1173.126	11.192

From Table 3.8, it is easy to see that the overall model quality for the top 10 models of the HsHsp90-MEEVD:HsHopTPR2AB:HsHsp70-PTIEEVD complex is acceptable (N-DOPE Z< -0.5). Minimisation of the HsTPR2ab2yeast model 13 returns the lower Cα-RMSD, N-DOPE Z and Rosetta Energy scores, at the cost of moving an equal amount of residues from the "favoured" to the "allowed" region, and *vice versa* in the Ramachandran plot (Figure 3.22).

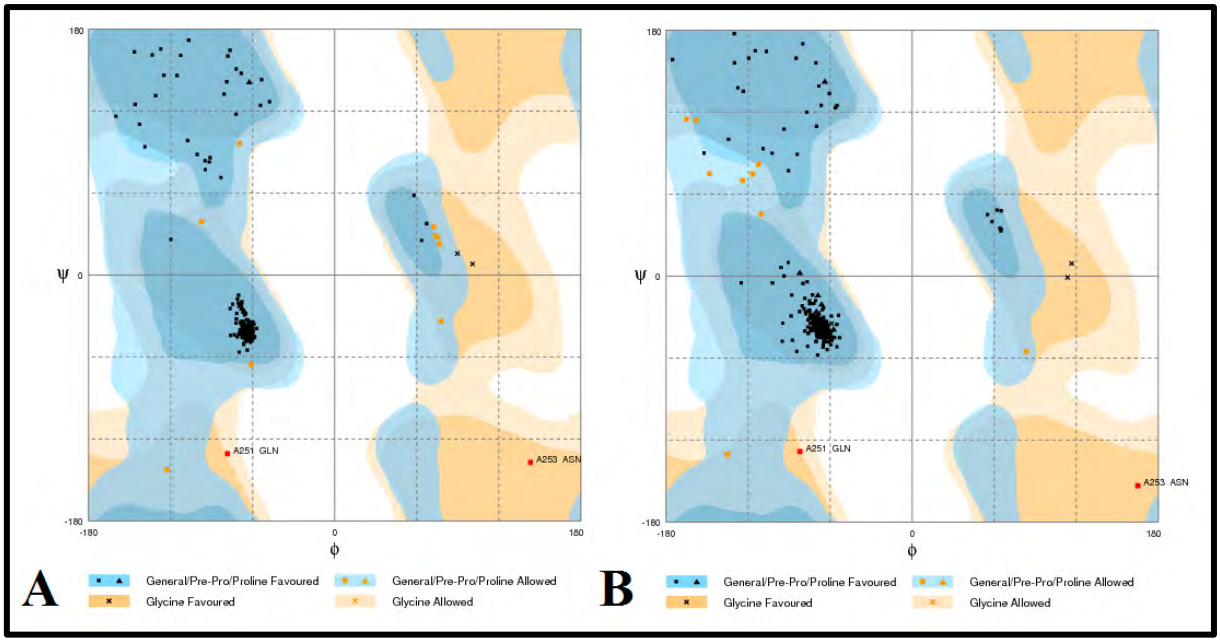


Figure 3.22: Ramachandran plots for HsTPR2ab model 13. A) general plot before minimisation. B) General plot after minimisation. Red squares indicate residues (GLN251 and ASN253) occupying disallowed regions. Black triangles and squares represent amino acids in the "Favoured" regions, Orange triangles and squares represent amino acids in the "allowed" regions.

This is a small trade-off for a better quality model and hence, the minimised model would preferably have been used for further analysis. MetaMQAPII analysis indicates that there are a lot of regions in the model that appear to degrade with minimisation and there are very few regions that display improvement (Figure 3.23).

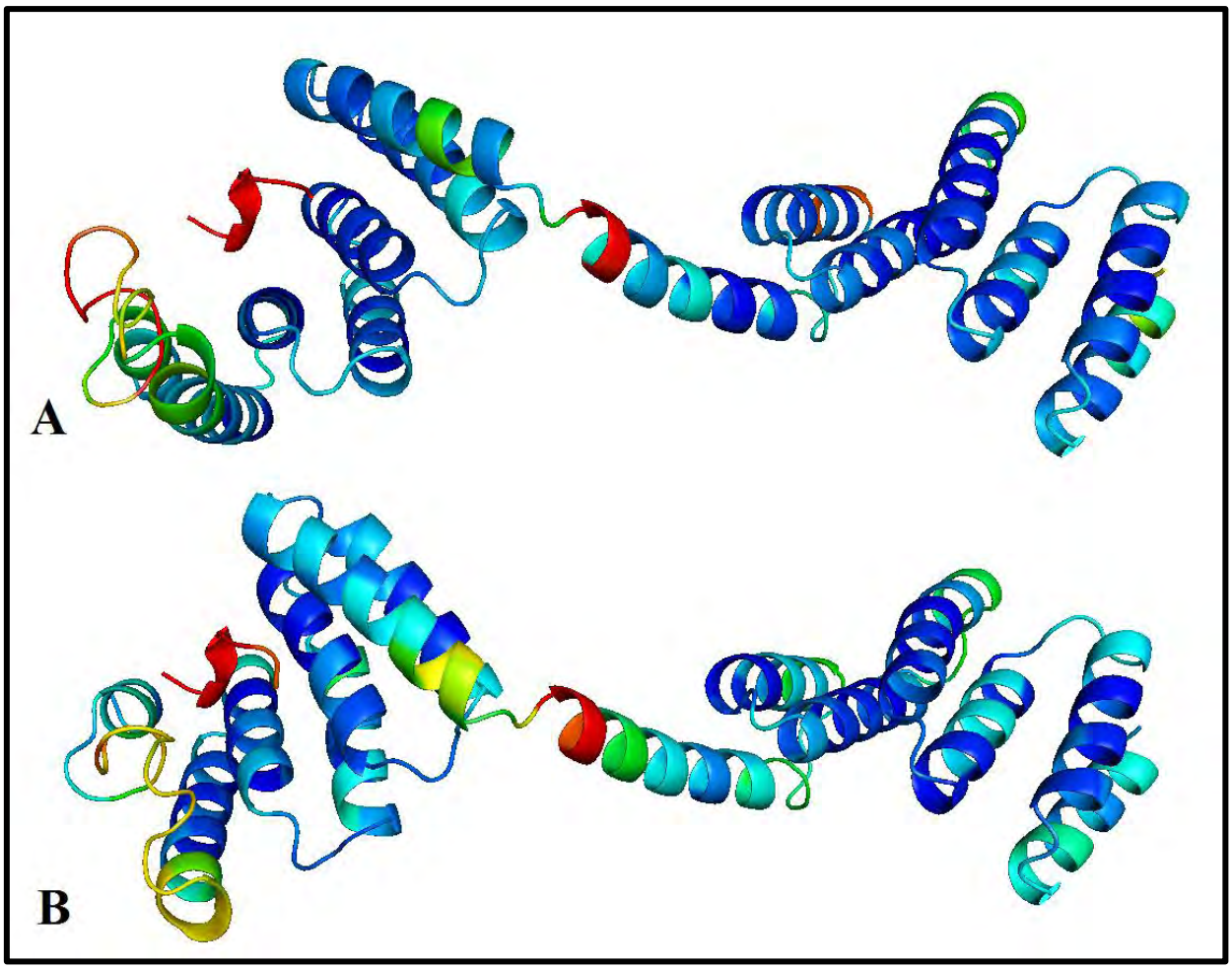


Figure 3.23: A) MetaMQAPII rendition of chain A and C HsTPR2ab model 13. B) MetaMQAPII rendition of minimised HsTPR2ab2yeast model 13 chain A and C.

Unfortunately, in spite of increased overall model quality, the minimisation step disrupted an important feature of the TPR2 domain: the interactions between the three residues of the REY clamp (TYR130, GLU161 and ARG165). Both PIC intra-protein analysis (Table 3.11) and PyMOL (see Figure 3.24) failed to detect sufficient interactions between these three residues in the minimised model. This is an indication that for this region of the model, model quality is poor.

Table 3.11: Intra-protein interactions calculated by the PIC webserver pertaining to the residues within the REY clamp in HsTPR2ab2yeast model 13. The final column represents interactions between any of the three REY residues with any non-REY residue in the structure.

	Main-Side chain Interactions	Side-Side chain Interactions	Ionic Interactions	Cation-Pi Interactions	Non intra-REY interactions
HsTPR2ab-2yeast model 13	0	2 x TYR-GLU 3 x ARG-TYR	1 x GLU-ARG	1 x TYR-ARG	1 x Cation-pi TYR-ARG127
Minimised	0	0	1 x GLU-ARG	0	1 x Main-Side ALA129-TYR

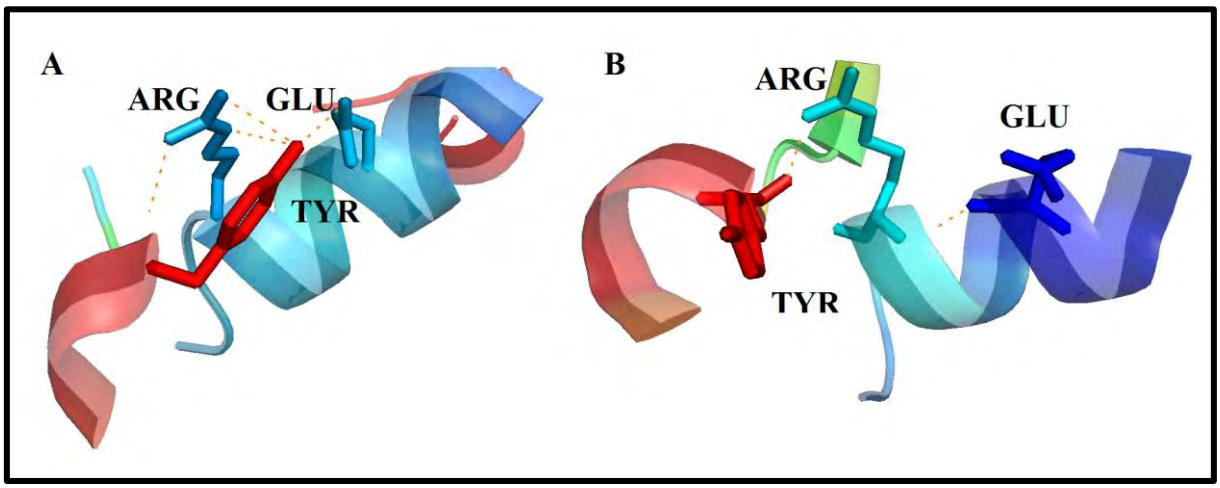


Figure 3.24: The REY clamp residues (stick representation; TYR130, GLU161 and ARG165) in chain A of HsTPR2ab model 13. A) MetaMQAPII rendition before minimisation. B) MetaMQAPII rendition after minimisation. The orange dashes indicate polar contacts predicted between the three residues.

In the next few pages, the assessment results of the best of the Hsp90:HopTPR2A&B:Hsp70 complex models in *P. falciparum* Hop are discussed.

From Table 3.12, below, it is easy to see that the overall model quality for the top 10 models ofPfHsp90-MEEVD:PfHopTPR2AB:PfHsp70-PTVEEVDis acceptable (N-DOPE Z< -0.5). The Cα-RMSD score is increased after minimisation, reducing the backbone similarity of the model to the template.

Table 3.12: Energy scores for top 10 PfHsp90-MEEVD:PfHopTPR2AB:PfHsp70-PTVEEVD complex models. Models are listed in order of decreasing DOPE-Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
Template1: 3UPV_mod.pdb	-2.176	-130.692	7.582
Template 2: 3uq3_TPR2ab.pdb	-1.473	-212.989	0.000
Minimised_PfTPR2ab2yeast.B99990014_1.pdb	-1.383	-717.198	7.376
PfTPR2ab2yeast.B99990014.pdb	-1.056	1248.952	7.180
PfTPR2ab2yeast.B99990022.pdb	-1.052	1324.550	7.236
PfTPR2ab2yeast.B99990098.pdb	-1.042	1180.747	7.146
PfTPR2ab2yeast.B99990027.pdb	-1.028	1039.171	7.210
PfTPR2ab2yeast.B99990018.pdb	-1.025	1222.645	7.414
PfTPR2ab2yeast.B99990011.pdb	-1.012	1265.783	7.155
PfTPR2ab2yeast.B99990090.pdb	-1.011	1495.175	7.463
PfTPR2ab2yeast.B99990017.pdb	-1.003	1287.585	7.537
PfTPR2ab2yeast.B99990046.pdb	-0.993	1371.101	7.237
PfTPR2ab2yeast.B99990032.pdb	-0.986	1612.909	7.536
PfTPR2ab2yeast.B99990031.pdb	-0.985	1221.850	7.120

Refinement through minimisation of the PfTPR2ab2yeast Model 14 returns the lower N-DOPE Z and Rosetta energy scores, and restores several residues from the "favoured" to the "allowed" regions, and the single residue in the disallowed (or "outlier") to the "allowed" region in the Ramachandran plot (Figure 3.25). The Cα-RMSD score is increased after minimisation, reducing the backbone similarity of the model to the template.

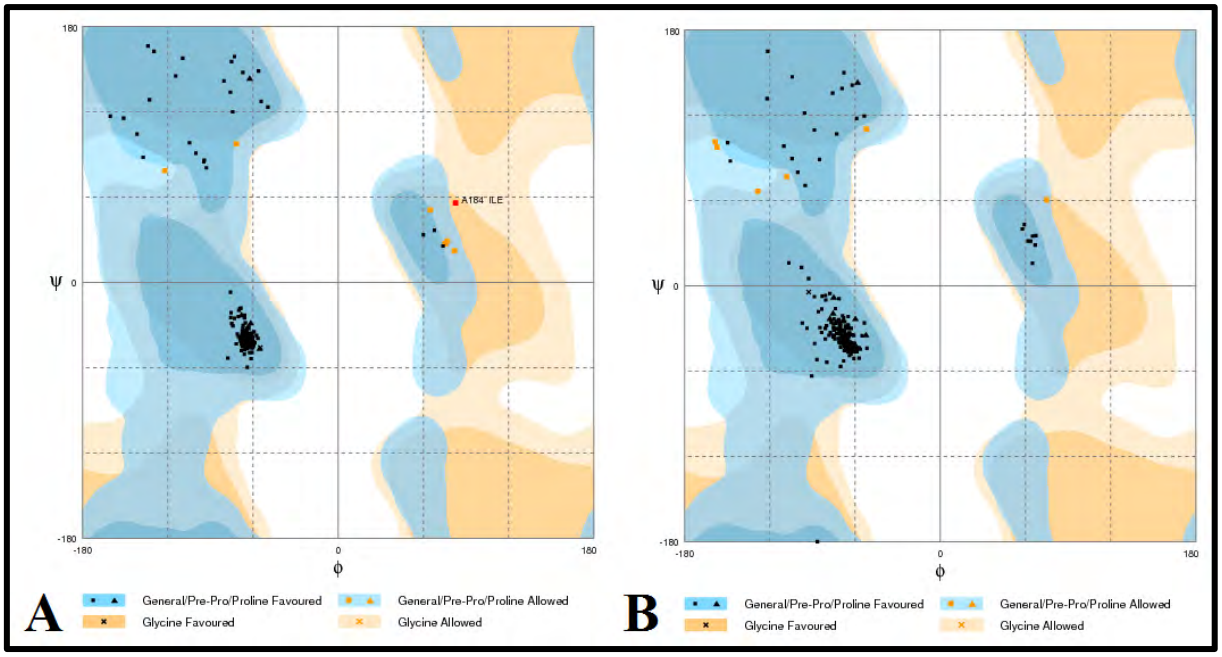


Figure 3.25: Ramachandran plots for PfTPR2ab2yeast model 14. A) General plot before minimisation. Red square indicates residue ILE181 occupying a disallowed region B) General plot after minimisation. Black triangles and squares represent amino acids in the "Favoured" regions, Orange triangles and squares represent amino acids in the "allowed" regions.

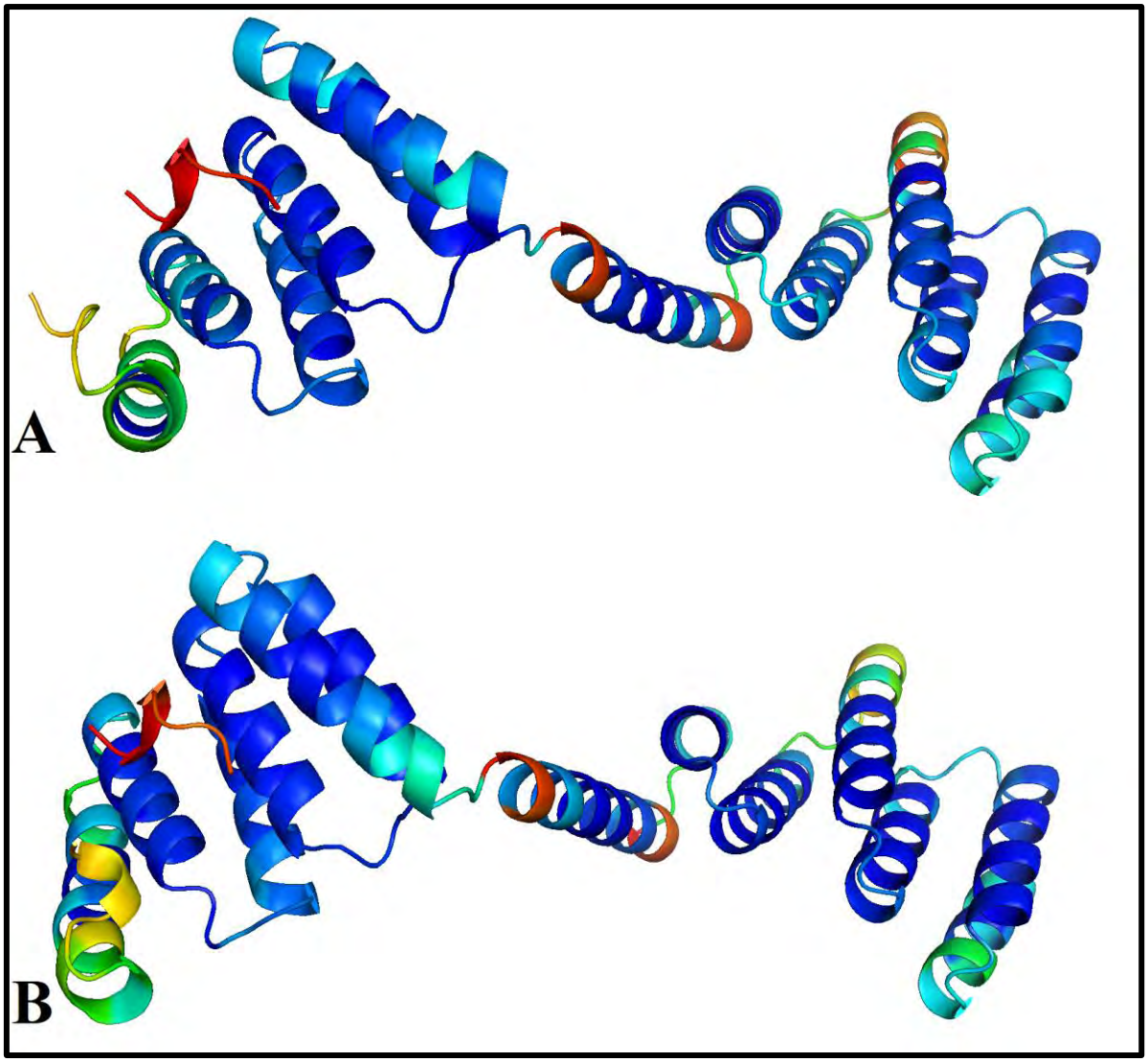


Figure 3.26: A) MetaMQAPII rendition of chain A and C PfTPR2ab2yeast model 14. B) MetaMQAPII rendition of minimised PfTPR2ab model 14 chain A and C.

Additionally, in spite of increased overall model quality, the minimisation step again disrupted the interactions between the three residues of the REY clamp (TYR130, GLU161 and ARG165). Both PIC intra-protein analysis (Table 3.13) and PyMOL (see Figure 3.27) failed to detect sufficient interactions between these three residues in the minimised model. This is an indication that for this region of the model, model quality is poor. This is too large a trade-off for a better quality model and hence, the original model was used for further analysis.

Table 3.13: Intra-protein interactions calculated by the PIC webserver pertaining to the residues within the REY clamp PfTPR2ab2yeast model 14. The final column represents interactions between any of the three REY residues with any non-REY residue in the structure.

	Main-Side chain Interactions	Side-Side chain Interactions	Ionic Interactions	Cation-Pi Interactions	Non intra-REY Interactions
PfTPR2ab-2yeast model 14	2 x ARG-TYR	2 x TYR-GLU 2 x ARG-TYR 3 x ARG-GLU	1 x GLU-ARG	1 x TYR-ARG	3 x Ionic 1 x GLU-ARG164 1 x ASP132 - ARG 1 x HIS139 - GLU
Minimised	2 x ARG-TYR	2 x ARG-GLU	1 x GLU-ARG	0	4 x Side-Side 3 x ARG164-GLU 1 x ARG-HIS139 3 x Ionic 1 x GLU-ARG164 1 x ASP132 - ARG 1 x HIS139 - GLU

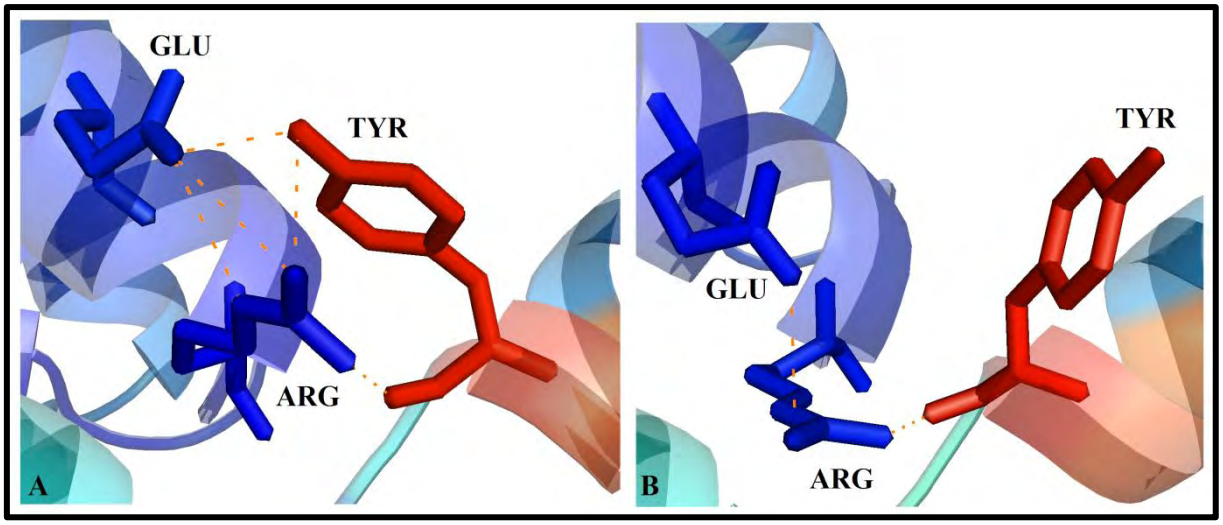


Figure 3.27: The REY clamp residues (stick representation; TYR130, GLU161 and ARG165) in chain A of PfTPR2ab model 14. A) MetaMQAPII rendition before minimisation. B) MetaMQAPII rendition after minimisation. The orange dashes indicate polar contacts predicted between the three residues.

In the next few pages, the assessment results for the best of the Hsp90:HopTPR2 complex models in human Hop are discussed.

Table 3.14: Energy scores for top 10 HsHsp90M&Cdomains:HsHopTPR2 complex. Models are listed in order of decreasing DOPE-Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
tailess_schmidCYS.pdb	-1.069	-161.600	0.000
MinimisedHscomplex_07_1.pdb	-1.095	-1600.531	5.850
Hs_complex_single.B99990007.pdb	-0.668	2950.236	2.648
Hs_complex_single.B99990029.pdb	-0.658	3464.726	2.648
Hs_complex_single.B99990084.pdb	-0.649	3002.862	2.650
Hs_complex_single.B99990085.pdb	-0.640	3058.365	2.652
Hs_complex_single.B99990074.pdb	-0.628	3003.778	2.650
Hs_complex_single.B99990027.pdb	-0.623	3157.533	2.627
Hs_complex_single.B99990080.pdb	-0.606	3117.747	2.641
Hs_complex_single.B99990054.pdb	-0.605	3465.792	2.653
Hs_complex_single.B99990006.pdb	-0.604	3114.043	2.632
Hs_complex_single.B99990008.pdb	-0.597	3529.762	2.625

The overall model quality for the top 10 HsHsp90M&Cdomains:HsHopTPR2 complex models was acceptable (N-DOPE Z < -0.5). For general analysis of the HsHsp90M&Cdomains:HsHopTPR2 complex model 07, it was plain to see from Table 3.14 that minimisation resulted in drastic improvement of the N-DOPE Z and Rosetta energy scores, however, the Cα-RMSD score was also increased.

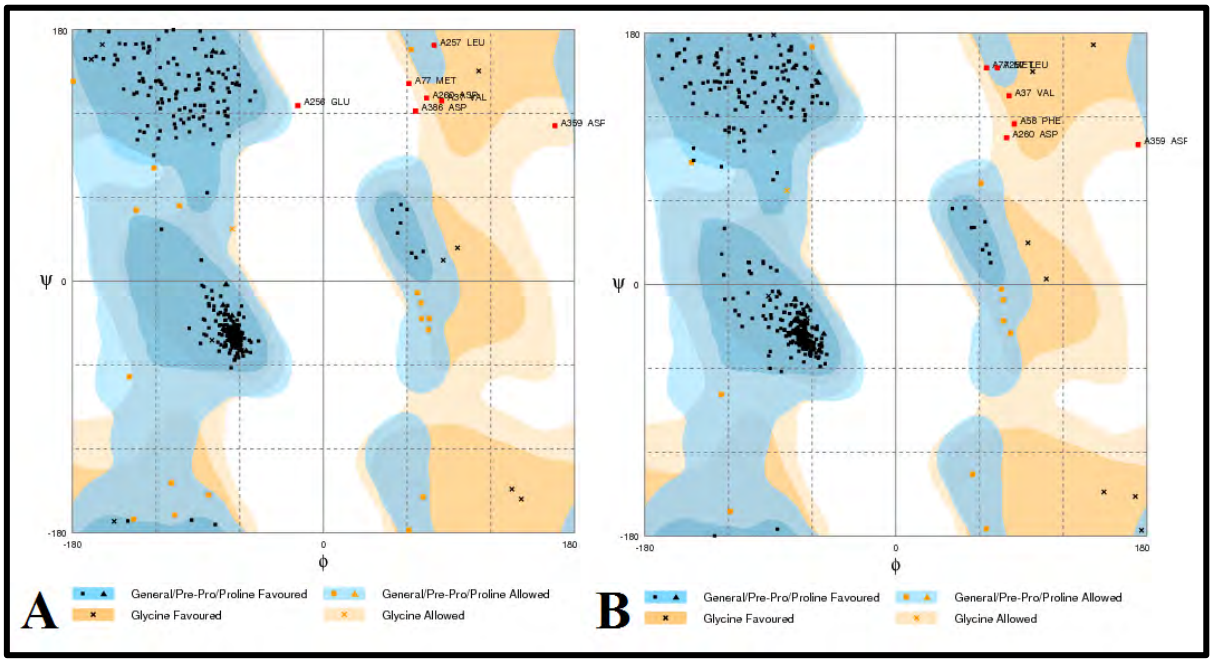


Figure 3.28: Ramachandran plots for the Hsp90 half of HsComplex model 07. A) General plot before minimisation. Red squares indicate residues (VAL37, MET77, GLU256, LEU257, ASP260, ASP359 and ASP386) occupying disallowed regions. B) General plot after minimisation. Red squares indicate residues (VAL37, PHE58, MET77, LEU257, ASP260 and ASP359) occupying disallowed regions. Black triangles and squares represent amino acids in the "Favoured" regions, Orange. triangles and squares represent amino acids in the "allowed" regions.

Based on MetaMQAPII scores, the Hsp90 half of the complex showed overall improvement with minimisation (see Figure 3.29), which was confirmed by the Ramachandran plots for the original and minimised versions of Hsp90 (see Figure 3.28). Minimisation reduced the number of residues occupying both the disallowed and allowed phi and psi regions, and increased the percentage of residues within the favoured region from 93.7% to 95.5%.

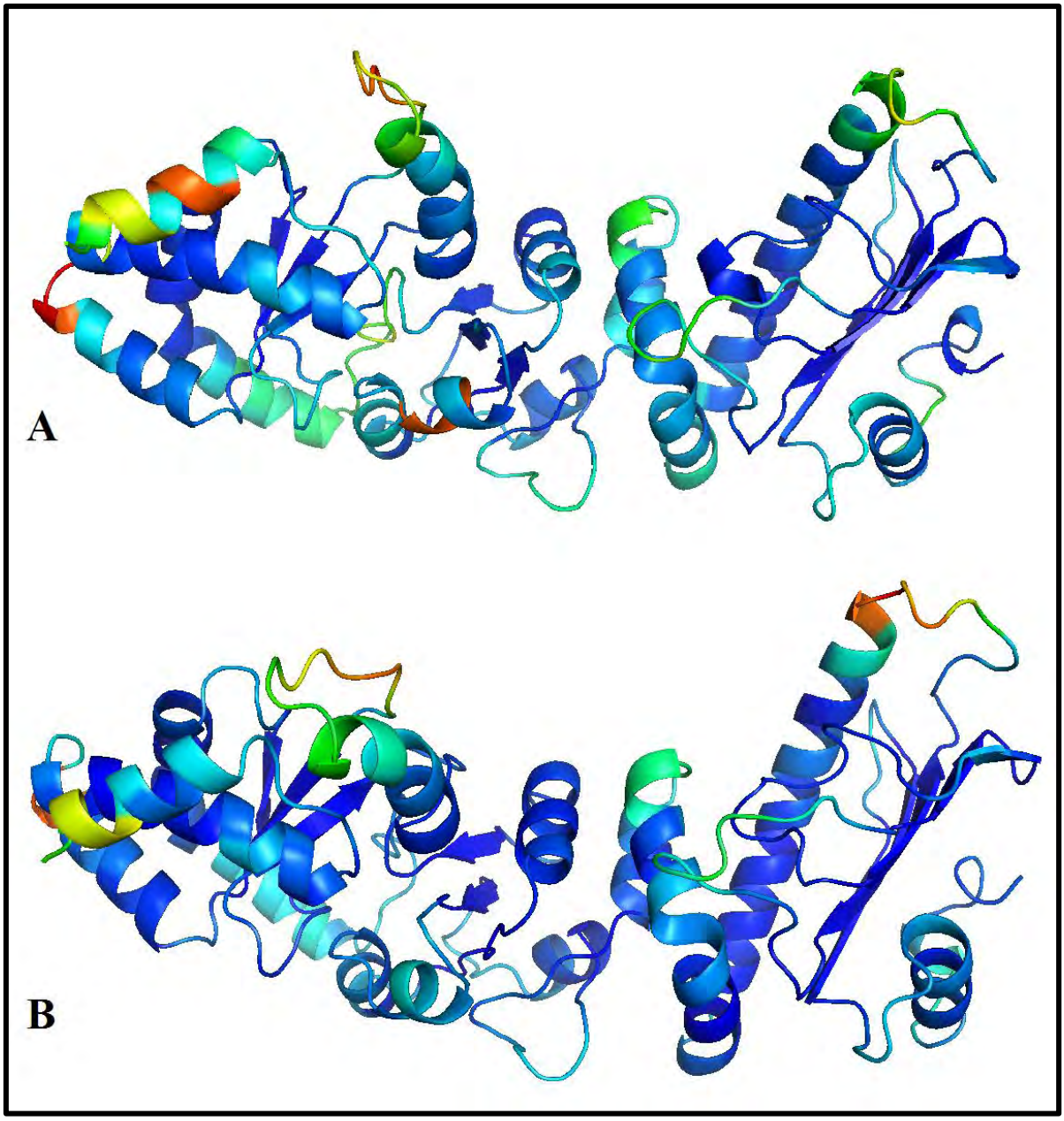


Figure 3.29: A) MetaMQAPII rendition of chain A HsComplex model 07. B) MetaMQAPII rendition of minimised chain A HsComplex model 7.

Based on MetaMQAPII representation, the Hop half of the complex appeared to decrease slightly in quality with minimisation (see Figure 3.29), while Ramachandran plots indicated that the minimised version of Hop was of better quality (see Figure 3.30). Minimisation halved the number of residues occupying the allowed phi and psi regions, and increased the percentage of residues within the favoured region from 95.8% to 97.9%. This is almost the expected value (98%) for a native model.

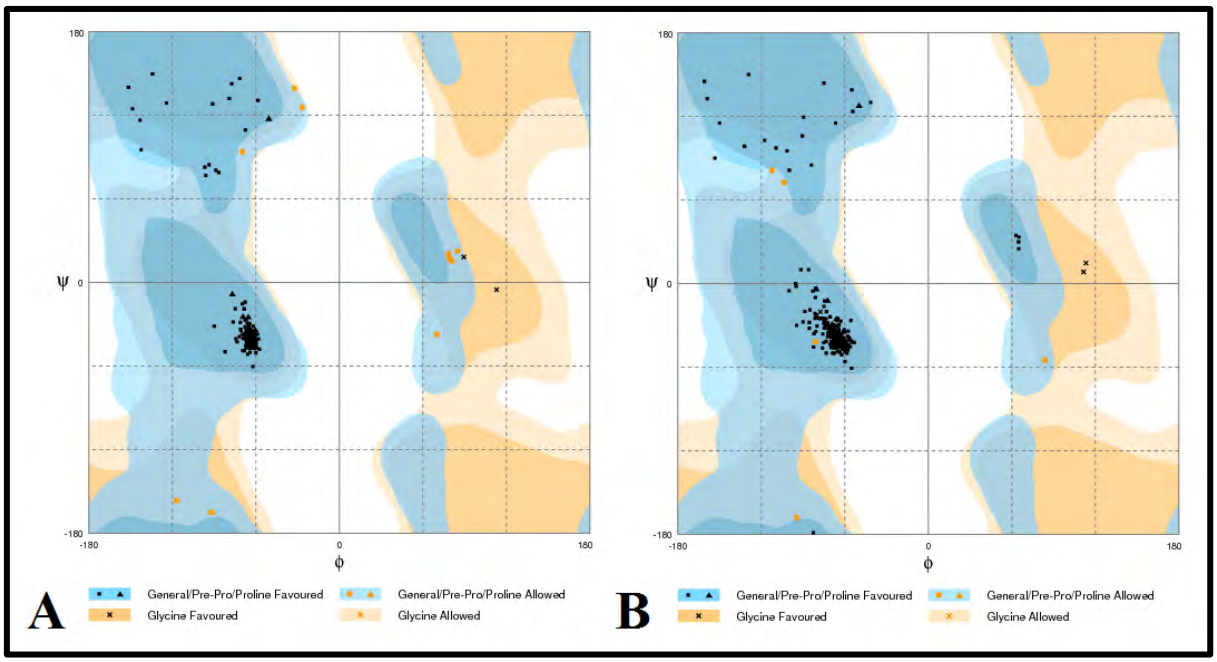


Figure 3.30: Ramachandran plots for the Hop half of HsComplex model 07. A) General plot before minimisation. B) General plot after minimisation. Black triangles and squares represent amino acids in the "Favoured" regions, Orange triangles and squares represent amino acids in the "allowed" regions.

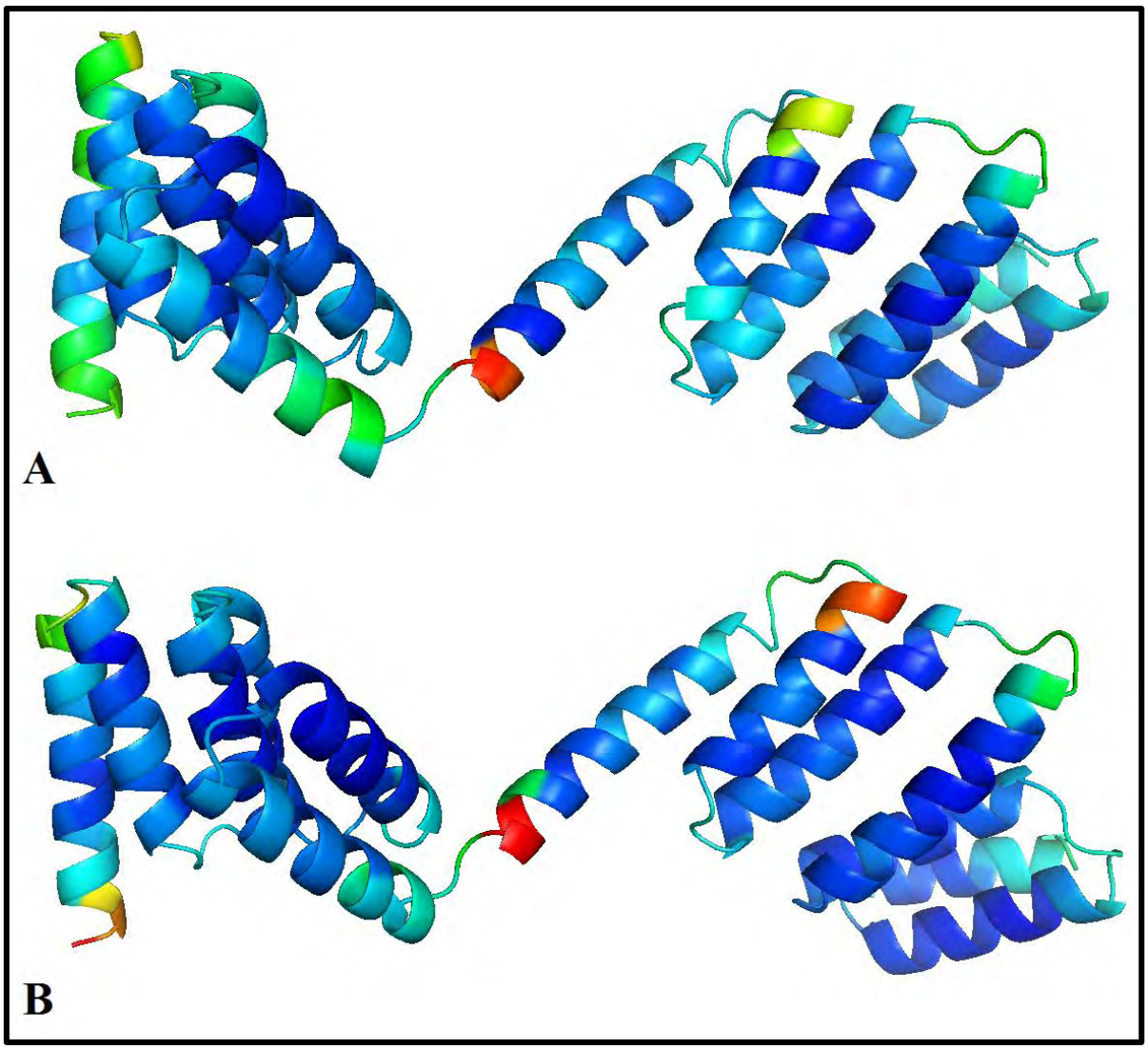


Figure 3.31: A) MetaMQAPII rendition of chain B HsComplex model 07. B) MetaMQAPII rendition of minimised chain B HsComplex model 07.

Table 3.15: Intra-protein interactions calculated by the PIC webserver pertaining to the residues with the REY clamp within the Hop half ofHsTPR2: HsHsp90 model 14. The final column represents interactions between any of the three REY residues with any non-REY residue in the structure.

	Main-Side chain Interactions	Side-Side chain Interactions	Ionic Interactions	Cation-Pi Interactions	Non intra-REY interactions
Hop half of HsTPR2: HsHsp90 model 14	2 x ARG-TYR	2 x TYR-GLU 2 x ARG-TYR 3 x ARG-GLU	1 x GLU-ARG	1 x TYR-ARG	1 x Ionic 1 x GLU-LYS562 1 x Cation-Pi 1 x TYR-LYS562
Minimised	1 x ARG-TYR	2 x TYR-GLU 3 x ARG-TYR	2 x GLU-ARG	1 x TYR-ARG	2 x M-S: 2 x ARG-LEU533 3 x S-S: 1 x LYS562-GLU 2 x ARG-GLU537 3 x Ionic: 1 x GLU536-ARG 1 x GLU537-ARG 1 x GLU-LYS562

PyMOL and PIC analysis of the REY clamp showed almost no degradation of interactions between the three residues (see Figure 3.32 and Table 3.15). Therefore, based on the conservation of the orientation of the tyrosine residue, as well as the cation-π interactions and several hydrogen bonds, the minimised model was of preferable quality and was used for subsequent studies on protein-protein interactions in Chapter 4.

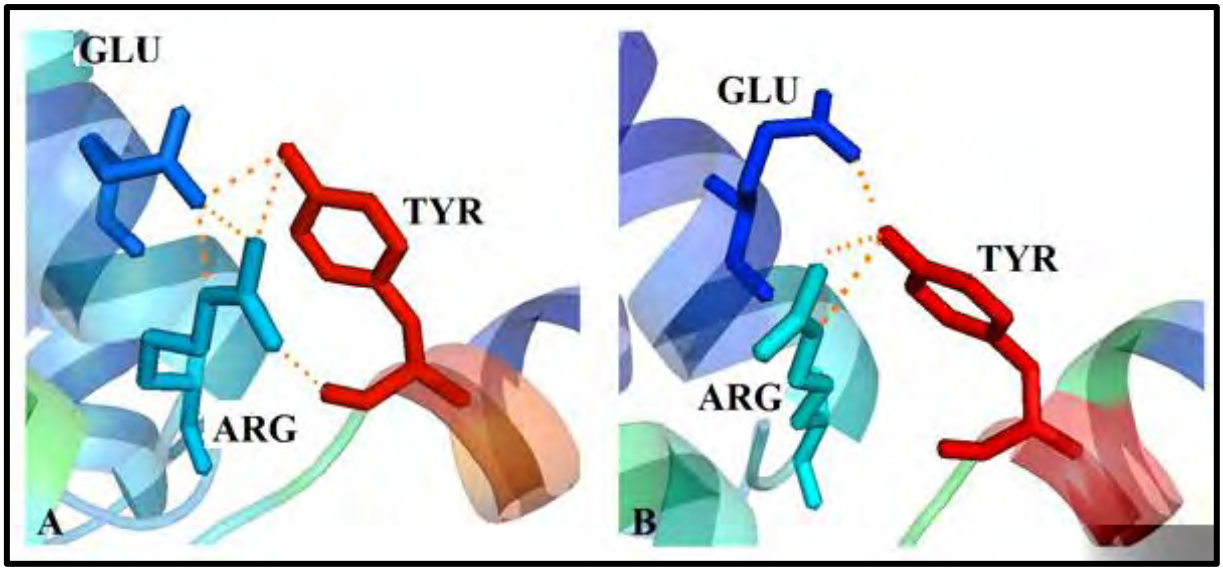


Figure 3.32: The REY clamp residues (stick representation) in chain B HsComplex model 7. A) MetaMQAPII rendition before minimisation. B) MetaMQAPII rendition after minimisation. The orange dashes indicate polar contacts predicted between the three residues.

In the next few pages, the assessment results for the best of the Hsp90:HopTPR2 complex models in *P.falciparum* Hop are discussed.

Table 3.16: Energy scores for top 10 PfHsp90M&Cdomains:PfHopTPR2 complex models. Models are listed in order of decreasing DOPE-Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
Template 1: 4GCO_mod	-2.168	-95.861	N/A
Template 2: tailess_schmidCYS.pdb	-1.069	-161.600	0.000
MinimisedPfMulti_102.pdb	-0.824	-1625.737	15.991
Pfmulti.B99990102.pdb	-0.423	4140.058	15.627
Pfmulti.B99990077.pdb	-0.409	4228.389	15.636
Pfmulti.B99990056.pdb	-0.392	4259.119	15.618
Pfmulti.B99990059.pdb	-0.389	3754.028	15.612
Pfmulti.B99990036.pdb	-0.375	4770.000	15.577
Pfmulti.B99990075.pdb	-0.353	4427.120	15.595
Pfmulti.B99990019.pdb	-0.346	4512.396	15.614
Pfmulti.B99990063.pdb	-0.343	4042.399	15.587
Pfmulti.B99990041.pdb	-0.342	4174.044	15.604

For general analysis of the PfHsp90M&Cdomains:PfHopTPR2 complex model 102, it was plain to see from Table 3.16 that minimisation resulted in drastic improvement of the DOPE-Z and Rosetta energy scores, while the Cα-RMSD score was only increased by a small fraction.

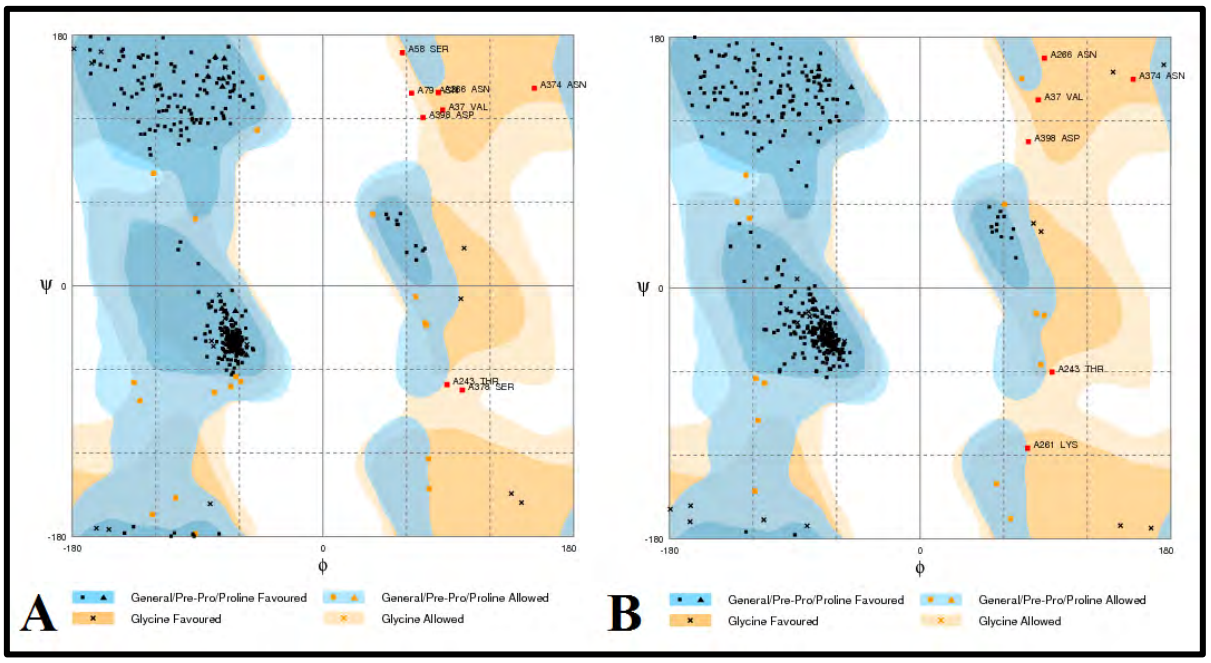


Figure 3.33: Ramachandran plots for the Hsp90 half of PfComplex model 102. A) General plot before minimisation. Red squares indicate residues (VAL37, ASN79, THR243, ASN266, ASN374, SER378 and ASP398) occupying disallowed regions. B) General plot after minimisation. Red squares indicate residues (VAL37, THR243, LYS261, ASN266, ASN374 and ASP398) occupying disallowed regions. Black triangles and squares represent amino acids in the "Favoured" regions, Orange triangles and squares represent amino acids in the "allowed" regions.

Ramachandran plots for the original and minimised versions of Hsp90 (see Figure 3.33) showed that minimisation reduced the number of residues occupying both the disallowed and allowed phi and psi regions, and increased the percentage of residues within the favoured region from 93.4% to 95.1%. Based on MetaMQAPII scores, the Hsp90 half of the complex showed overall improvement with minimisation (see Figure 3.34).

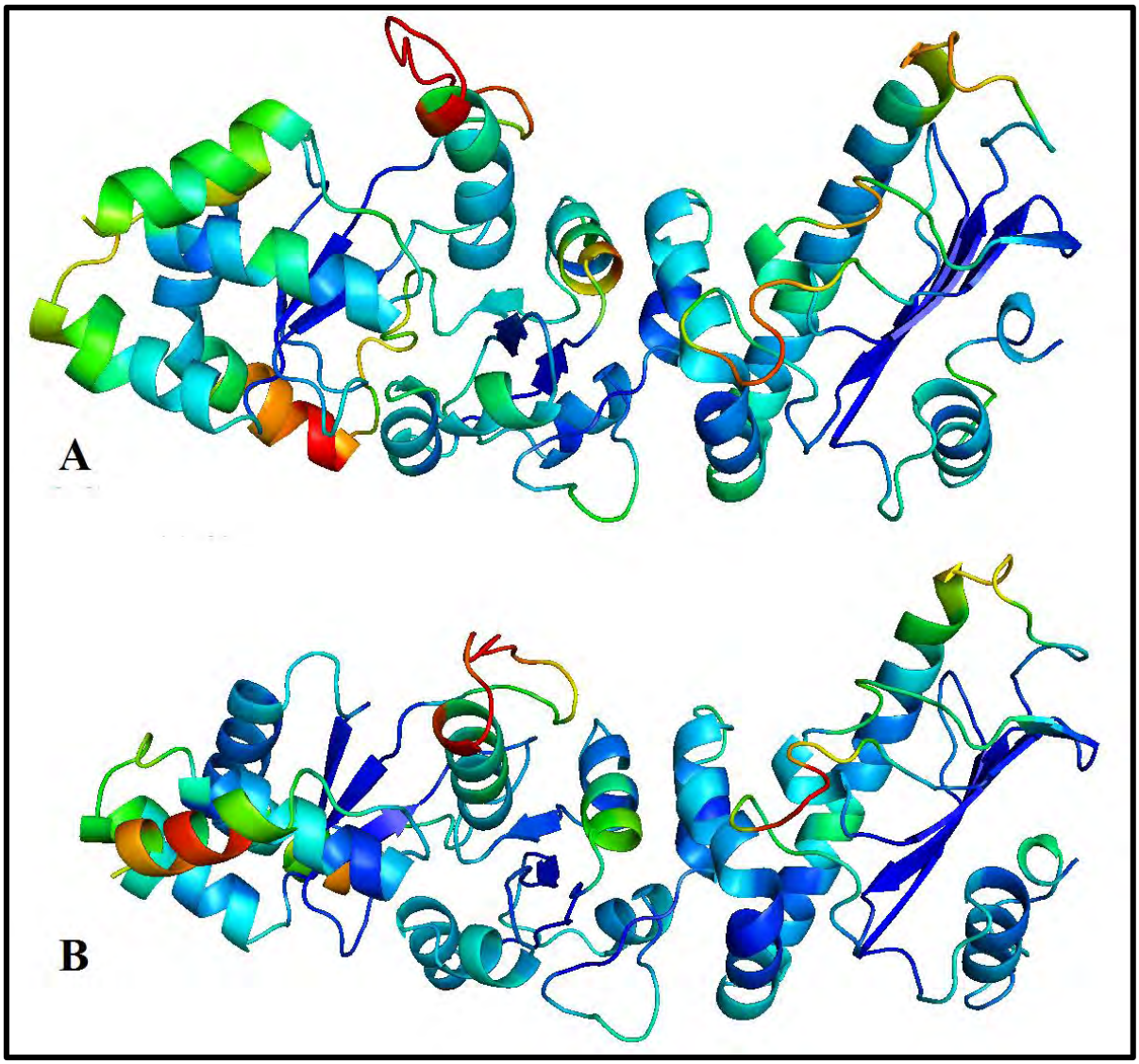


Figure 3.34: A) MetaMQAPII rendition of chain A Pfmulti model 102. B) MetaMQAPII rendition of minimised chain A Pfmulti model 102.

Based on MetaMQAPII representation, the Hop half of the complex appeared to improve slightly in quality with minimisation (see Figure 3.35), which was confirmed with Ramachandran plots indicating that the minimised version of Hop is of better quality (see Figure 3.36).

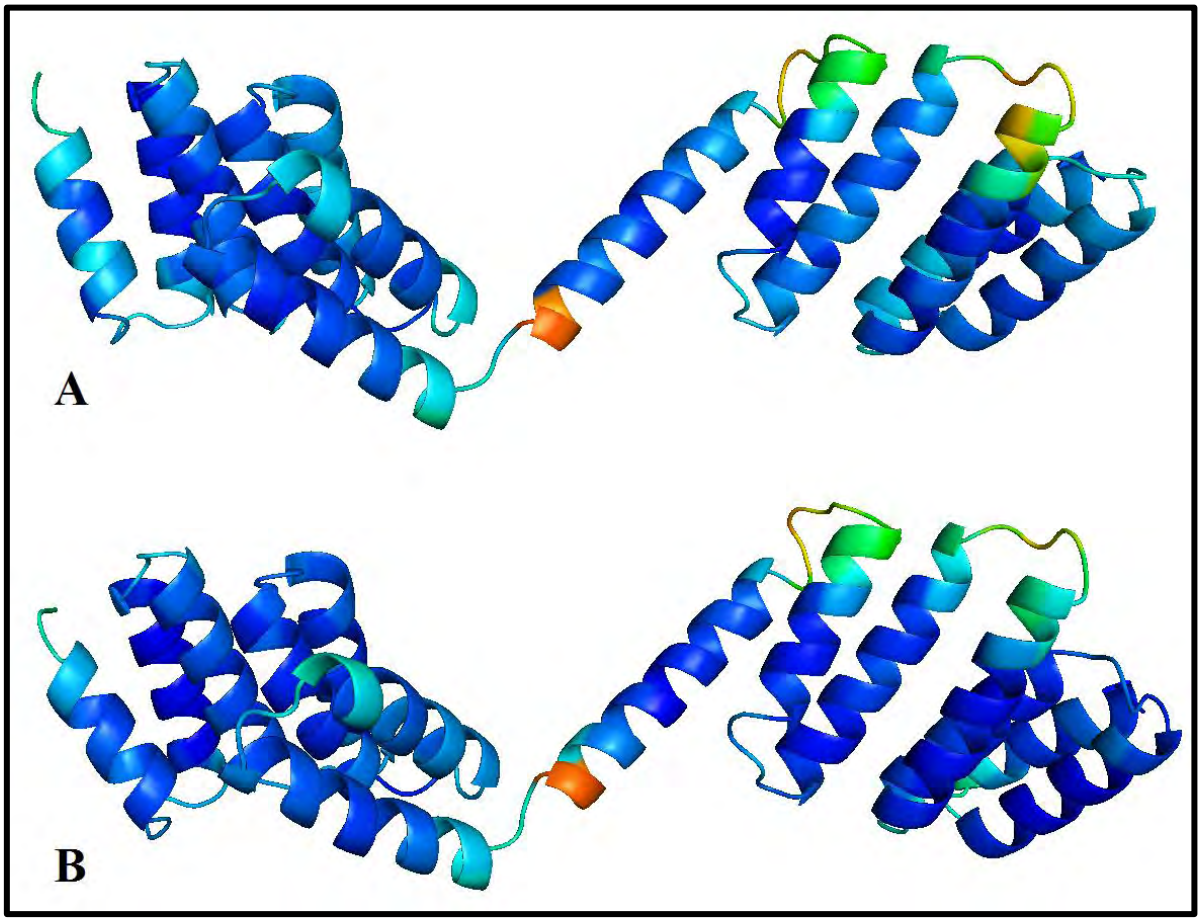


Figure 3.35: A) MetaMQAPII rendition of chain B Pfmulti model 102 B) MetaMQAPII rendition of minimised chain B Pfmulti model 102.

Minimisation moved the single residue in the disallowed region to its preferred region and dramatically reduced the number of residues occupying the allowed phi and psi regions, which increased the percentage of residues within the favoured region from 94.4% to 98.8%. This is greater than the expected value (98%) for a native model.

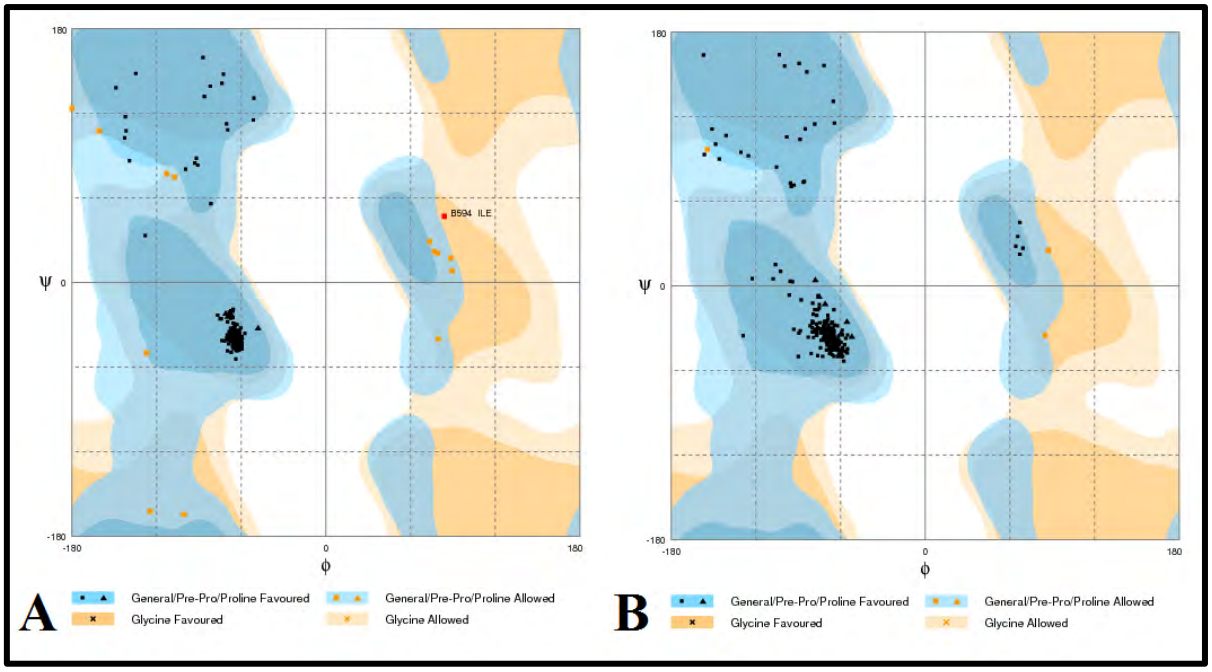


Figure 3.36: Ramachandran plots for the Hop half of Pfmulti model 102. A) General plot before minimisation. . Red square indicates residue ILE594 occupying a disallowed region. B) General plot after minimisation. Black triangles and squares represent amino acids in the "Favoured" regions, Orange triangles and squares represent amino acids in the "allowed" regions.

PyMOL and PIC analysis of the REY clamp showed a small degree of degradation in interaction between the three residues, and again a change in orientation of the tyrosine (see Figure 3.37 and Table 3.17). However, based on the conservation of the Cation-π interaction and several hydrogen bonds, this was a small trade-off; the minimised model was of preferable quality and was used for subsequent studies on protein-protein interactions in Chapter 4.

Table 3.17: Intra-protein interactions calculated by the PIC webserver pertaining to the

residues with the REY clamp. The final column represents interactions between any of the three REY residues with any non-REY residue in the structure.

	Main-Side chain Interactions	Side-Side chain Interactions	Ionic Interactions	Cation-Pi Interactions	Non intra-REY interactions
Hop half of PfTPR2: PfHsp90 model 102	2 x ARG-TYR	2 x TYR-GLU 2 x ARG-TYR 1 x ARG-GLU	1 x GLU-ARG	1 x TYR-ARG	1 x S-S: 1 x ARG-HIS549 3 x Ionic: 1 x ASP542-ARG 1 x GLU-ARG574 1 x HIS549-GLU
Minimised	1 x ARG-TYR	2 x ARG-GLU	2 x GLU-ARG	1 x TYR-ARG	1 x Ionic 1 x HIS549-GLU

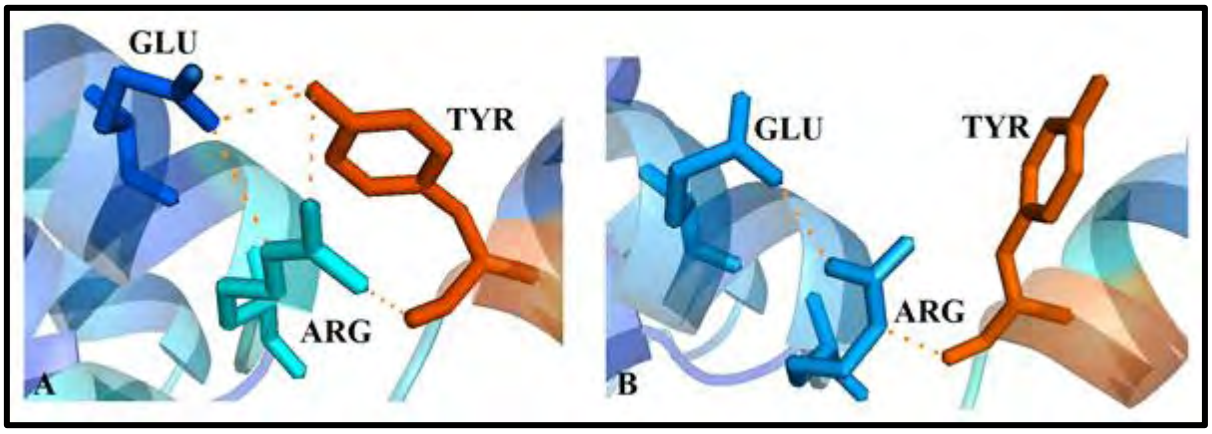


Figure 3.37: The REY clamp residues (stick representation) in chain B Pfmulti model 102. A) MetaMQAPII rendition before minimisation. B) MetaMQAPII rendition after minimisation. The orange dashes indicate polar contacts predicted between the three residues.

3.3.3.3 Model Analysis for DP Structures

The models built for the DP structures are validated in the same way as the complex models, but are only analysed and compared in brief. This is primarily because there is very little understanding on the role of these domains in Hop functioning.

Table 3.18: Energy scores for top 10 HsDP1 models. Models are listed in order of decreasing N-DOPE Z scores. Models are listed in order of decreasing DOPE-Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
Minimised_2LLV_average_1.pdb	-2.279	-140.029	0.000
Minimised_ HsDP1.B99990080_1.pdb	-1.765	650.570	9.489
HsDP1.B99990080.pdb	-1.623	485.168	9.265
HsDP1.B99990094.pdb	-1.546	663.021	9.271
HsDP1.B99990084.pdb	-1.499	634.739	9.234
HsDP1.B99990031.pdb	-1.494	706.112	9.457
HsDP1.B99990062.pdb	-1.490	641.174	9.214
HsDP1.B99990036.pdb	-1.488	717.452	9.213
HsDP1.B99990008.pdb	-1.487	654.019	9.217
HsDP1.B99990073.pdb	-1.461	811.782	9.498
HsDP1.B99990022.pdb	-1.436	617.179	9.261
HsDP1.B99990089.pdb	-1.425	620.153	9.406

The overall model quality for the top 10 HsHopDP1 models was acceptable (N-DOPE Z < -1.0). For general analysis of the HsHopDP1 model 80, it was plain to see from Table 3.18 that minimisation resulted in only minor improvement of the N-DOPE Z and Rosetta energy scores; however, the Cα-RMSD score was increased. Based on MetaMQAPII scores, the model showed overall degradation with minimisation (see Figure 3.38), which was confirmed by the Ramachandran plots for the original and minimised versions of HsHopDP1 model 80 (see Figure 3.39). Minimisation reduced the number of residues occupying allowed phi and psi regions, and increased the percentage of residues within the favoured region from 90.7% to 92.0%.

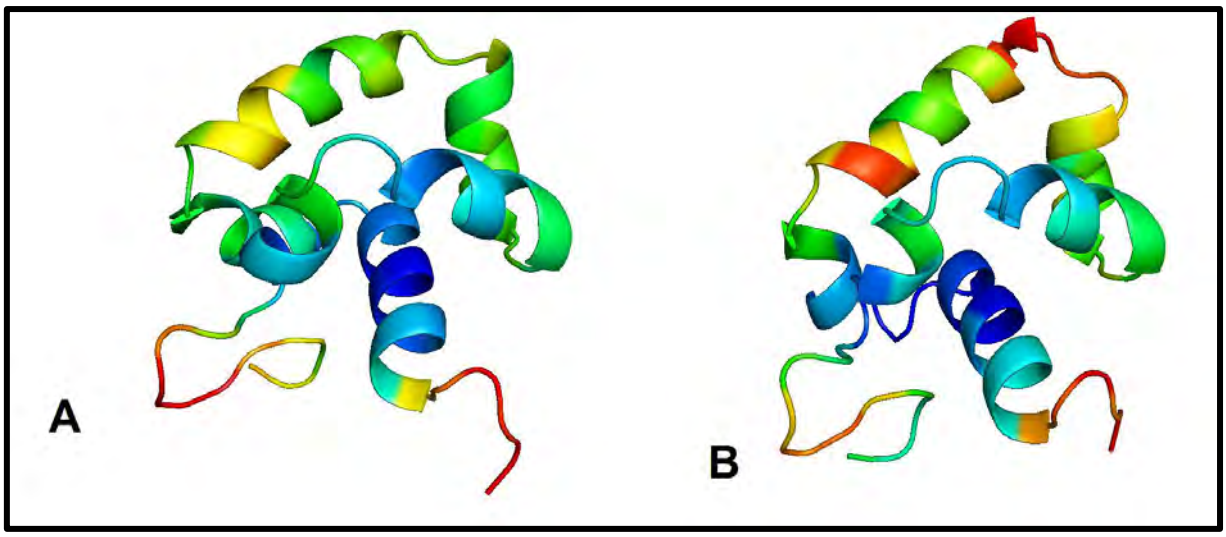


Figure 3.38: A) MetaMQAPII rendition of HsDP1 model 80 B) MetaMQAPII rendition of minimised HsDP1 model 80.

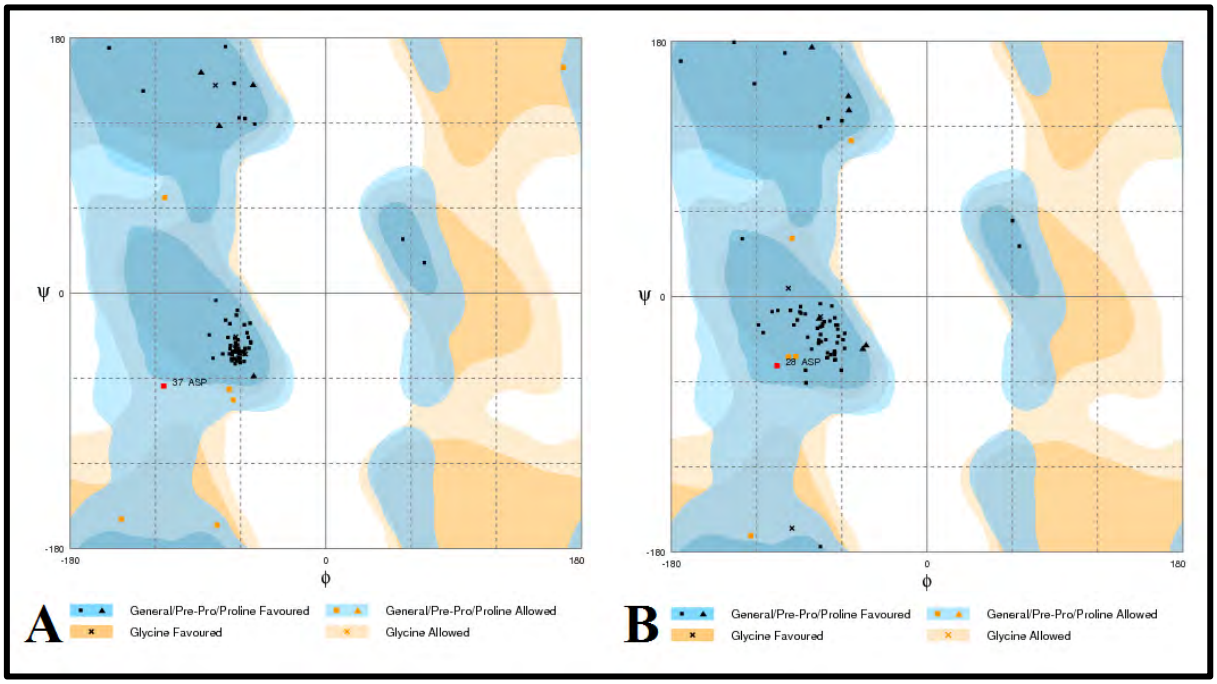


Figure 3.39: Ramachandran plots for HsDP1 model 80. A) General plot before minimisation. Red square indicates residue ASP37 occupying a disallowed region. B) General plot after minimisation. Red square indicates residue ASP28 occupying a disallowed region. Black triangles and squares represent amino acids in the "Favoured" regions, Orange triangles and squares represent amino acids in the "allowed" regions.

Table 3.19: Energy scores for top 10 PfDP1 models. Models are listed in order of decreasing N-DOPE Z scores. Models are listed in order of decreasing DOPE-Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
Minimised_2LLV_average_1.pdb	-2.279	-140.029	0.000
Minimised_PfDP1.B99990066_1.pdb	-1.140	-140.555	12.356
PfDP1.B99990066.pdb	-0.707	398.247	11.675
PfDP1.B99990016.pdb	-0.665	615.473	11.349
PfDP1.B99990024.pdb	-0.631	623.253	11.441
PfDP1.B99990082.pdb	-0.628	632.107	11.183
PfDP1.B99990007.pdb	-0.617	640.611	11.236
PfDP1.B99990094.pdb	-0.614	352.044	11.537
PfDP1.B99990061.pdb	-0.614	364.981	11.632
PfDP1.B99990003.pdb	-0.608	333.015	11.357
PfDP1.B99990048.pdb	-0.597	452.358	11.255
PfDP1.B99990018.pdb	-0.595	333.345	11.276

The overall model quality for the top 10 PfHopDP1 models was acceptable (N-DOPE Z < -5.0). For general analysis of the PfHopDP1 model 66, it was plain to see from Table 3.19 that minimisation resulted in dramatic improvement of the N-DOPE Z and Rosetta energy scores; however, the Cα-RMSD score was increased. Based on MetaMQAPII scores, the model showed overall improvement with minimisation (see Figure 3.40), which was confirmed by the Ramachandran plots of the original and minimised versions of Hsp90 (see Figure 3.41). Minimisation moved the single residue in the disallowed region to its preferred region, reduced the number of residues occupying allowed phi and psi regions, and increased the percentage of residues within the favoured region.

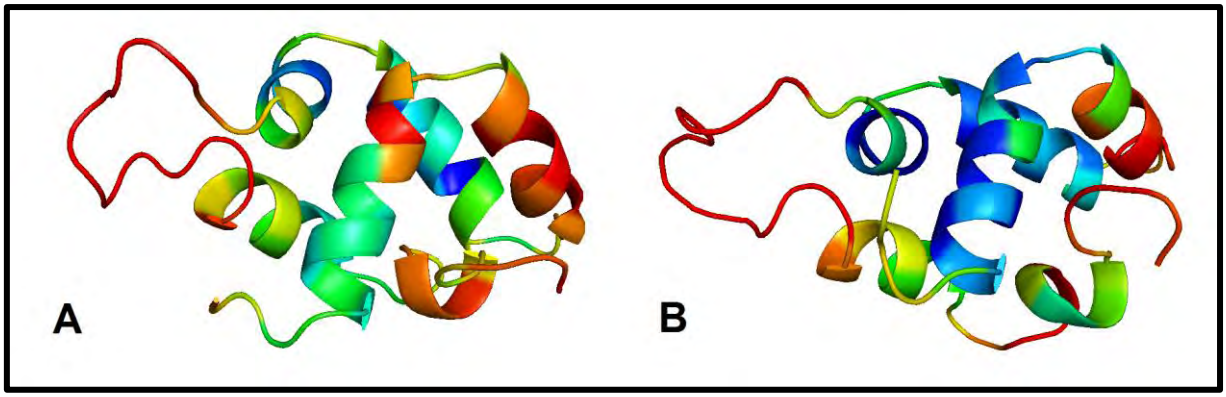


Figure 3.40: A) MetaMQAPII rendition of PfDP1 model 66 B) MetaMQAPII rendition of minimised PfDP1 model 66.

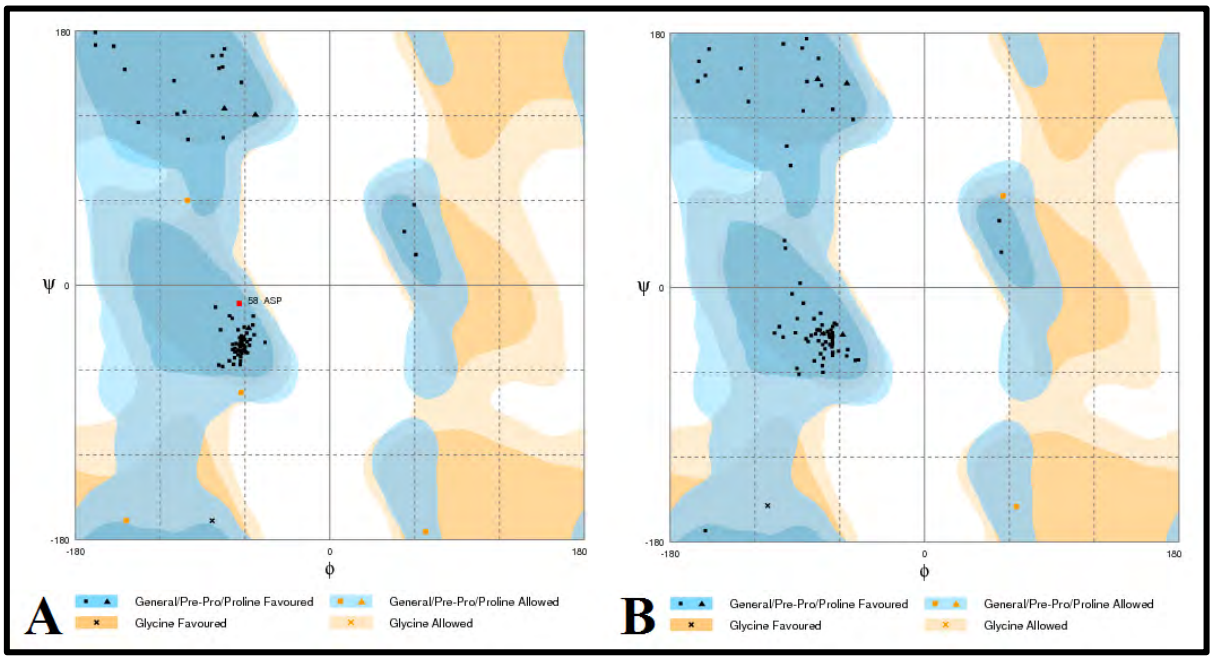


Figure 3.41: Ramachandran plots for PfDP1 model 66. A) General plot before minimisation. Red square indicates residue ASP58 occupying a disallowed region. B) General plot after minimisation. Black triangles and squares represent amino acids in the "Favoured" regions, Orange triangles and squares represent amino acids in the "allowed" regions.

DP1 is only clearly defined by conserved motifs in fungal, plant and certain invertebrate species, as was discussed in Chapter 2. Mammalian and monotreme Hop sequences are so well conserved it was difficult to discern whether DP1 was conserved as a single domain, but meme analysis identified at least the C-terminal end of DP2, TPR2A and the corresponding linker between the two domains as a single motif (see Figure 2.6, Chapter 2). If Hop does indeed possess a recognisable DP1 domain, it is significantly different from that in ScHop,

which would explain why the HsHop models are of average quality. However, it is highly doubtful that protozoan species possess a recognisable DP1 domain, as was concluded in Chapter 2. This is most likely the reason why the PfDP1 models are of relatively low quality and these models must be treated with speculation as to their validity. If DP1 is conserved, it may only be conserved in the structural sense, which may be difficult to confirm with homology models.

Table 3.20: Energy scores for top 10 HsDP2 models. Models are listed in order of decreasing N-DOPE Z scores. Models are listed in order of decreasing DOPE-Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
2LLW1_state009.pdb	-2.166	83.419	0.000
Minimised_HsDP2.B99990072_1.pdb	-3.123	-164.081	5.623
HsDP2.B99990072.pdb	-2.302	485.168	5.822
HsDP2.B99990007.pdb	-2.286	663.021	5.899
HsDP2.B99990086.pdb	-2.271	634.739	5.817
HsDP2.B99990084.pdb	-2.251	706.112	5.860
HsDP2.B99990059.pdb	-2.249	641.174	5.831
HsDP2.B99990095.pdb	-2.238	717.452	5.847
HsDP2.B99990045.pdb	-2.219	654.019	5.825
HsDP2.B99990061.pdb	-2.217	811.782	5.871
HsDP2.B99990076.pdb	-2.214	617.179	5.801
HsDP2.B99990014.pdb	-2.211	620.153	5.844

The overall model quality for the top 10 HsHopDP2 models was excellent (N-DOPE Z <<< -1.0). For general analysis of the HsHopDP2 model 78, it was plain to see from Table 3.20 that minimisation resulted in dramatic improvement of the N-DOPE Z, Cα-RMSD and Rosetta energy scores. Based on MetaMQAPII scores, the model showed overall improvement with minimisation (see Figure 3.42), which was confirmed by the Ramachandran plots for the original and minimised versions of Hsp90 (see Figure 3.43). Minimisation moved the single residues in the disallowed and allowed regions to their preferred region, resulting in 100% of residues occupying their favoured regions.

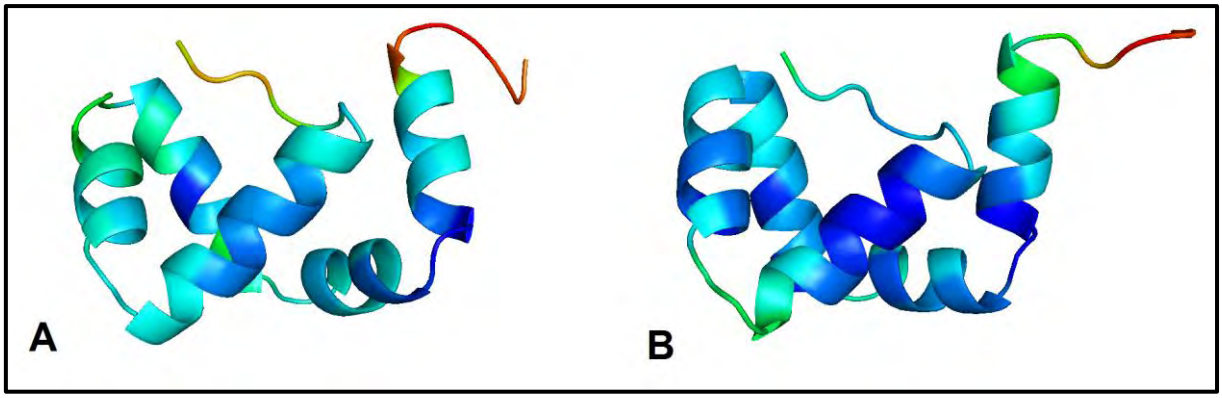


Figure 3.42: A) MetaMQAPII rendition of HsDP2 model 72 B) MetaMQAPII rendition of minimised HsDP2 model 72.

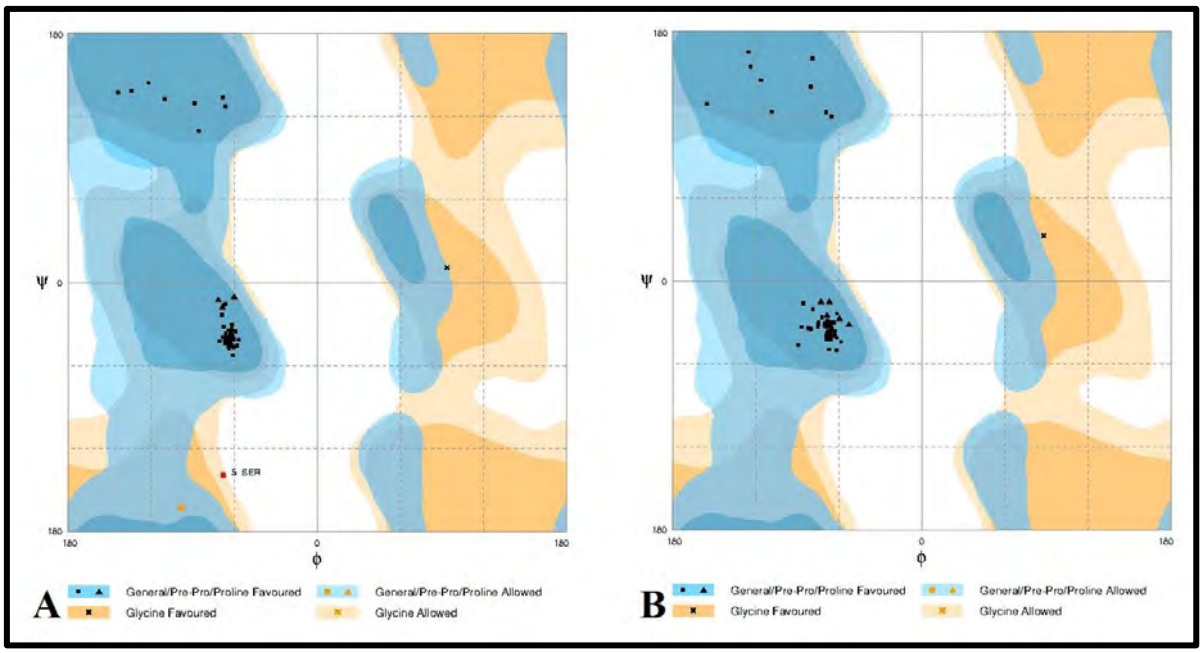


Figure 3.43: Ramachandran plots for HsDP2 model 78. A) General plot before minimisation. Red square indicates residue SER5 occupying a disallowed region. B) General plot after minimisation. Black triangles and squares represent amino acids in the "Favoured" regions, Orange triangles and squares represent amino acids in the "allowed" regions.

Table 3.21: Energy scores for top 10 PfDP2 models. Models are listed in order of decreasing N-DOPE Z scores. Models are listed in order of decreasing DOPE-Z scores.

Model	N-DOPE Z	Rosetta Energy	Cα-RMSD
2LLW1_state009.pdb	-2.166	83.419	0.000
Minimised_PfDP2009.B99990003_1.pdb	-2.341	-124.597	0.568
PfDP2009.B99990003.pdb	-2.029	446.246	0.490
PfDP2009.B99990070.pdb	-2.009	348.074	0.643
PfDP2009.B99990082.pdb	-1.972	382.159	0.586
PfDP2009.B99990013.pdb	-1.946	352.044	0.527
PfDP2009.B99990071.pdb	-1.939	364.981	0.674
PfDP2009.B99990047.pdb	-1.919	333.015	0.474
PfDP2009.B99990048.pdb	-1.918	452.358	0.858
PfDP2009.B99990073.pdb	-1.916	333.345	0.510
PfDP2009.B99990079.pdb	-1.913	327.039	0.839
PfDP2009.B99990028.pdb	-1.912	224.762	0.478

The overall model quality for the top 10 PfHopDP2 models was excellent (N-DOPE Z <<< -1.0). For general analysis of the PfHopDP2 model 03, it was plain to see from Table 3.21 that minimisation resulted in slight improvement of the N-DOPE Z and Rosetta energy scores; however, the Cα-RMSD score was increased. Based on MetaMQAPII scores, the model showed some degradation with minimisation (see Figure 3.42), which was contradicted by the Ramachandran plots for the original and minimised versions of PfHopDP2 model 03 (see Figure 3.43). Minimisation moved the single residue in the allowed regions to its preferred region, resulting in 100% of residues occupying their favoured regions.

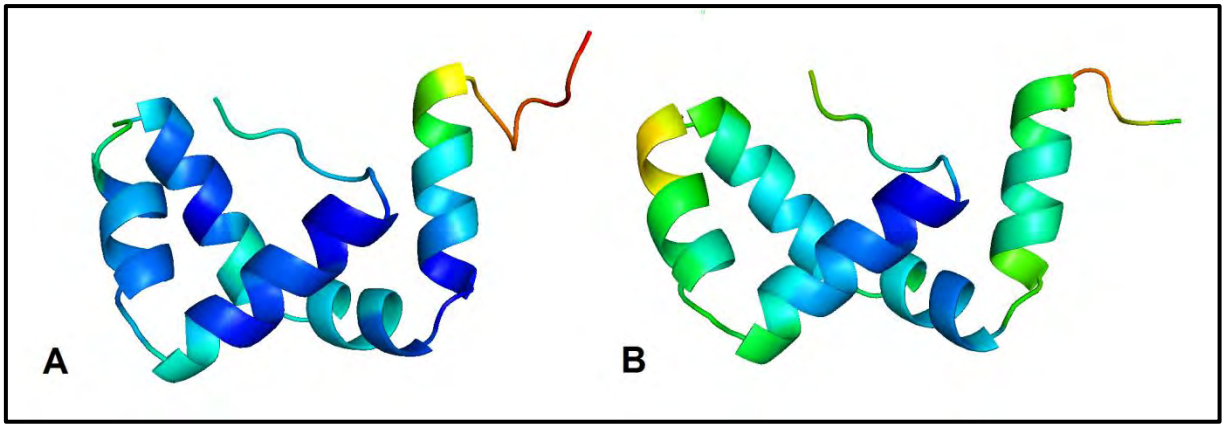


Figure 3.44: A) MetaMQAPII rendition of PfDP2009 model 03 B) MetaMQAPII rendition of minimised PfDP2009 model 03.

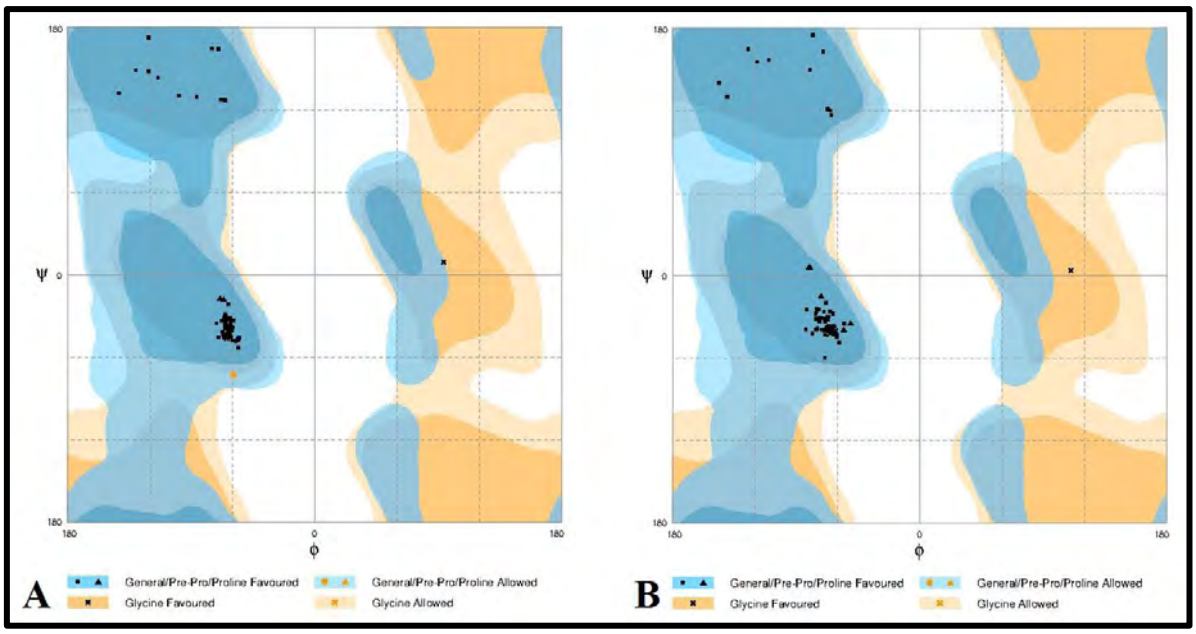


Figure 3.45: Ramachandran plots for PfDP2 model 03. A) General plot before minimisation. B) General plot after minimisation.

Based on the overall quality of all the DP models, it is clear that DP2 was far more accurately modelled than DP1, and in both cases the human models are superior to the *P. falciparum* ones. From Chapter 2, it was clear that DP2 is the most well conserved region in the Hop protein and may even be recognised as a conserved domain in other proteins, such as the XPC binding domain. This is most likely the reason for the production of good quality DP2 models for both species. The HsDP2 sequence shares higher identity with the yeast target than does PfDP2, resulting in better quality models for the human species.

3.4 Evaluation of Structures in Terms of Sequence Information

Both phylogenetic and MSA studies indicate that there are regions of similarity and difference in Hop across all eukaryote species analysed. This information was important for two reasons; firstly, human and *P. falciparum* models of the various domains were constructed based on yeast templates and it was important to take into account the differences in sequence and conserved domain organisation between templates and targets. To highlight these differences, the motif results for the three species of interest were aligned to their respective crystal structures, NMR structures and homology models in Figures 3.46 - 3.48.

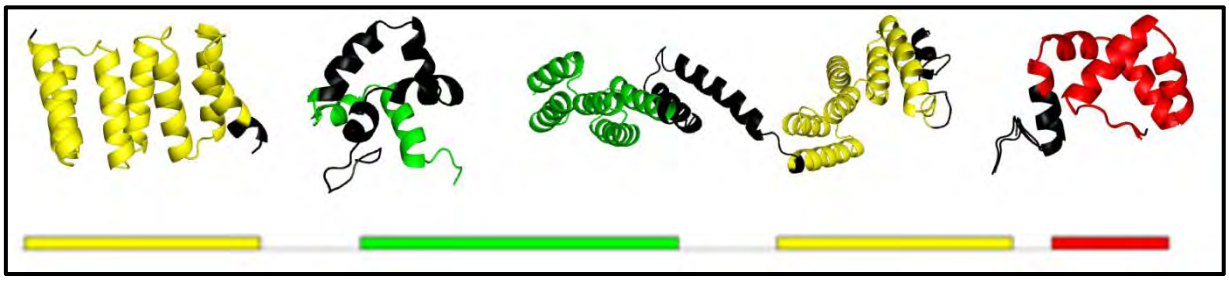


Figure 3.46: HsHop structures coloured and aligned to their corresponding motifs from Figure 2.6 (Chapter 2).

Figure 3.47: ScHop structures coloured and aligned to their corresponding motifs from Figure 2.6 (Chapter 2).

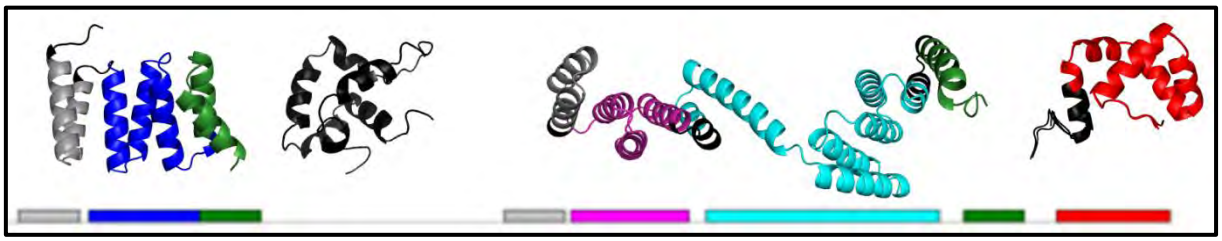


Figure 3.48: PfHop structures coloured and aligned to their corresponding motifs from Figure 2.6 (Chapter 2).

From these figures it seems that PfHop (Figure 3.48) is more similar to ScHop (Figure 3.47) with regards to domain organisation and that HsHop (Figure 3.46) is distinctly differently organised with respect to conserved domains. This motif structuring shared by vertebrates and

particularly within the mammals, does not necessarily mean that human hop is structured in a radically different way to Sc- and PfHop, and is most likely an artefact of the excellent conservation of Hop within the vertebrates, and MEME was simply positioning the motifs over the most well conserved regions within these species.

This is paradoxical, as when the template is pairwise aligned to the two targets, it turns out that ScHop shares higher identity with HsHop (Figure A6.1, Appendix 6) than with PfHop (Figure A6.2, Appendix 6) in spite of larger gaps in the alignment. As seen from Appendix 6, Figure A6.3, PfHop has higher identity to Hs- than ScHop. Similar scores were recorded from pBLAST of PfHop against the human genome to find HsHop (see Appendix 1).

3.5 Conclusions

It is clear from this chapter that there are some discrepancies in quality between HsHop and PfHop homology models for the various domains. This may be partly as a result of the fact that HsHop shares greater sequence identity with the template species than does PfHop. With the exception of the structure that was not published to the PDB ('tailess_schmidCYS'), all the templates selected were of good to excellent quality, and in most cases, shared at least 35% sequence identity with the targets. This was fortuitous, as there are relatively very few published structures for the various Hop domains, in complex or otherwise. However, while these templates are of good quality, it should not be taken for granted that structures published to the PDB (even those associated with publications) are without error and/or beyond improvement, as has been demonstrated through the "PDB Redo" project (Joosten, Joosten, Murshudov, & Perrakis, 2012). It was for this reason that all templates used were minimised, and both the minimised and original structures were analysed and validated as thoroughly as the homology models themselves. While the minimised templates were not used for homology modelling, this was an important step for understanding the potential negative or positive effects of minimisation of Hop structures.

Owing to time constraints, it was not possible to further validate the templates (those that are available) by being carefully compared to their redone versions in the PDBRedo database (Joosten et al., 2011). However, a brief scan of the templates currently available on PDBRedo (1ELW, 3UQ3 and 3UPV) shows that the basic statistics for those models, such as R-values, resolution and R-free values were identical or very close to those published within their respective PDB entries.

Overall, most homology models produced were of acceptable to very good quality. Homology models that were already of good quality appeared to be very minimally improved with minimisation, and in certain cases minimisation resulted in the degradation of unusual features, such as the interactions of the REY clamp. In other cases, such as for poorly scored homology models that were built on unsuitable or poor quality templates, minimisation appears to greatly improve the quality of these models, with minimal degradation of the REY clamp.

Chapter 4: Analysis of Binding Energies and Protein-protein Interactions

4.1 Introduction

Protein-protein interactions are the most important networks maintaining biological functions, from protein folding to programmed cell death, and represent a large and important class of targets for human therapeutics (Arkin & Wells, 2004; Ma & Nussinov, 2007). Viral and bacterial pathogens often rely on critical protein-protein interactions (such as the E1-E2 interactions in cervical cancer and fibronectin-binding proteins in bacterial infection) to allow infection (Arkin and Wells, 2004). Other biologically important processes are the result of unnatural protein-protein interactions that can cause disease, such as protein aggregations and malfunctioning voltage gated ion-channels in neurodegenerative diseases (Clare, Tate, Nobbs, & Romanos, 2000). Therefore, controlling protein-protein interactions has received increasing attention in drug target research.

Earlier this year, PfHop was localised and isolated from the trophozoite (infective) stage of the parasite, and characterised in depth, determining that PfHop potentially exists in association with PfHsp70 and PfHsp90 (Gitau et al., 2012). While these interactions are only thought to be possible in *P. falciparum*, there is still a lot of experimental work to be done to substantiate these interactions existence. The following section endeavours to describe, compare and contrast the differences between interactions in the several complexes predicted in the previous chapter. This work attempts to go some way toward identifying a means to design a selective drug that targets only parasite protein-protein interactions within the infected host.

4.1.1 Protein Interactions Calculator (PIC) and ROBETTA Alanine Scanning Webservice (AlaScan)

The PIC returns a number of interaction descriptions for several interaction types submitted per complex model(Tina, Bhadra, & Srinivasan, 2007). To cross-reference these results the complexed models were sent to AlaScan (Kortemme & Baker, 2002; Kortemme, Kim, & Baker, 2004). AlaScan, whose fundamental principles for calculating the effect on complex Rosetta energy by comparing the overall Rosetta energy for residue-by-residue alanine mutants, was used as the basis for the script detailed in the next section. All interactions reported for an interaction type for each model were summarised numerically in the tables in

this section. However, only residues that were detected by AlaScan to be important to complex formation were included (by residue identifiers) in the tables (e.g. see Table 4.1). In order to determine which interactions and important residues were conserved or analogous, the interacting residues were mapped to their respective homology models, which were aligned and viewed (see .pse files in the supplementary data) in PyMOL.

4.2 Methods and Software

4.2.1 Protein Interactions

Protein interactions for the various complexes were predicted using the PIC Webserver. These results were downloaded and visualised in PyMOL and Discovery Studio Visualiser and compared with interactions predicted by these programs. Further, residues that are important to complex formation were detected using ROBETTA Alanine Scanning Webserver. Of the residues presented in the tables, those that were not involved in analogous interactions (with respect to the model being compared) were highlighted in bold.

4.2.2 Interaction and Binding Energy Calculations

Binding and interaction energies for the various complexes were calculated both before and after minimisation. In the case of models comprising three proteins in complex (e.g. Hsp90:TPR2:Hsp70 complexes), both the binding energies for the overall complex and unbound (strictly between two proteins) interactions were calculated for each complex.

Both interaction (ΔI) and binding energies ($\Delta\Delta E$) were calculated with a script written for the purpose (See Appendix 6, Section D). Binding energies were calculated with the same standard score function used to calculate the Rosetta energy scores in PyRosetta based on a basic equation used by others (Kajander et al., 2009):

$$\Delta\Delta E = E(AB) - \Delta E(A) - \Delta E(B)$$

ΔE represents the energy required to convert an unbound proteinfrom its bound conformation to it solution-state conformation, i.e. A* to A as in Figure 4.1. The custom script assumes that simplyrepacking of the rotamers or minimisation of the entire protein is sufficient to convert an unbound proteinfrom its bound conformation to it solution-state conformation. The interaction energy between proteins A and B is the change in energy for the process of A* and B* forming a complex (energy changes as a result of the presence of the other partner):

$\Delta I = E(AB) - E(A) - E(B)$

Since the Rosetta score does not scale with any physical units and is not an actual physical energy, neither the physical interaction energy, nor the physical binding energy calculated by the script is an accurate energy value, however, the difference in scores was interpreted qualitatively. Greater scores indicate poorer binding/affinities, while lower scores indicate better binding/ affinities.

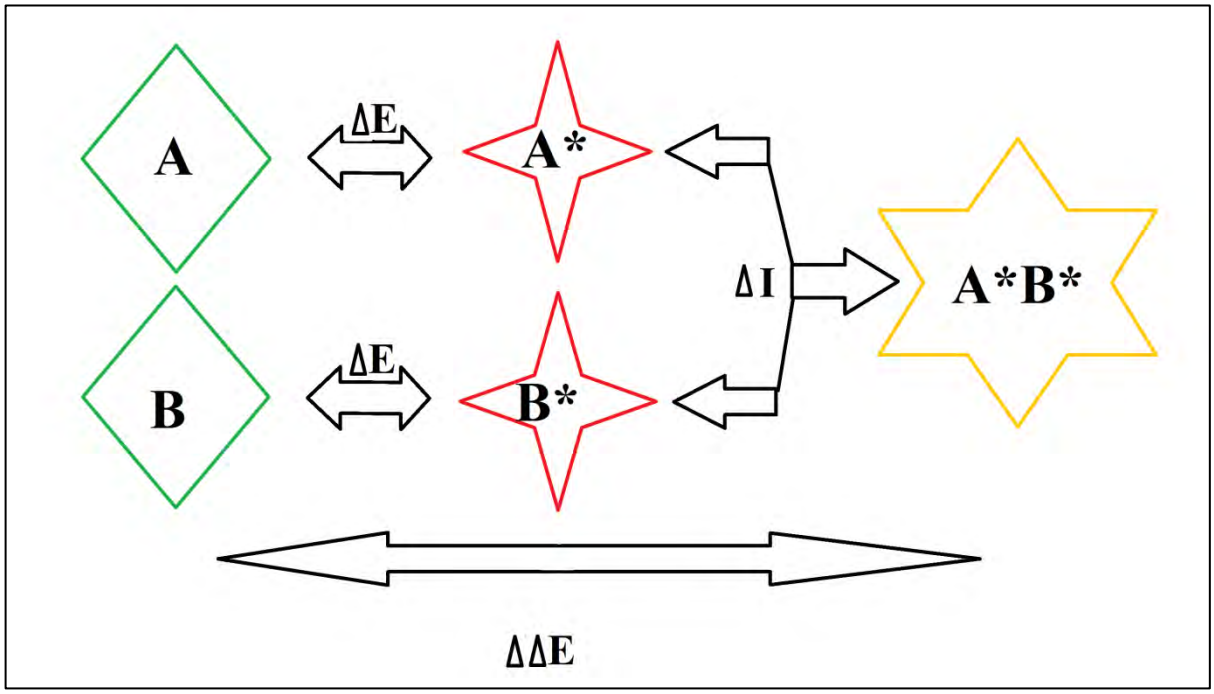


Figure 4.1: Simplified diagram illustrating the energies of interaction and binding involved in protein-protein interactions.

Another limitation of the approach detailed here, is that of statistical representation of the average binding energies. Ideally, most forms of refinement or minimisation are based on random sampling of a conformation space, they need to be iterative. With several rounds of minimisation, some feature of the model, such as overall energy score, converges at a global energy minimum. In other words, the minimised model ends up in a specific conformation more often than in others (this process has been implemented in more professionally used software and was discussed in clearer detail in Chapter 3, Sections 3.1.2 and 3.1.4). Unfortunately, owing to time limitation, this feature could not be incorporated into the script detailed in Appendix 2, Section D. Thus, all binding energy values reported in Section 4.5 of this chapter must be treated with caution as they are not representative of a mean value. This

limitation does not apply to interaction energy (no minimisation needs to take place to calculate interaction energy, as described previously); hence its inclusion in the study as a form of control.

4.3 Results & Discussion

4.3.1 Analyses of Templates

This section deals with interactions found within the templates used for homology modelling, and how minimisation affected or disrupted those interactions. This overview assisted with understanding both the advantages and disadvantages of minimisation.

Table 4.1: Comparison of the interacting residues in both the original and refined versions of 1ELW (i.e HsHopTPR1 complexed to HsHsc70-1 C-terminal peptide GPTIEEVD).

Model	Method	Hydrophobic interactions	Main-Side Chain Interactions	Side-Side Chain Hydrophobic Interactions	Ionic Interactions
Original	**PIC**	8	13	2	5
	Interacting residues found in AlaScan	A, LEU15, TYR27, PHE84. C; VAL8, ILE6	A; LYS8, ASN12, ASN43, ***LYS73***, ARG77. C; ASP12, GLU10, GLU9	A; LYS73. C; ASP12	A; LYS73, ARG77. C; ASP12, GLU10, GLU9
Minimised	**PIC**	9	13	3	5
	Interacting residues found in AlaScan	A, LEU15, TYR27, PHE84. C; VAL8, ILE6	A; LYS8, ASN12, ASN43, ARG77. C; ASP12, GLU10, GLU9	A; ***LYS50***, LYS73, ***SER76***. C; ASP12, ***GLU10***, ***GLU9***	A; ***LYS50***, LYS73, ARG77. C; ASP12, GLU10, GLU9

From Table 4.1, it's clear to see that the minimised structure displayed more interactions than the original. However it is only the additional side-chain to side-chain hydrophobic interaction that was detected by alanine scanning. It also appears that LYS50 replaces the ARG77 for its role in one of the ionic interactions. The original analysis of this template described the "dicarboxylatebinding clamp" residues as make the largest favourable contributions to the stability of both peptide complexes (Kajander et al., 2009). The binding residues that were reported to contribute most significantly are LYS8, LYS73, and ASN43 as well as one conserved arginine, ARG77 as was found for both the original and minimised versions of this template. Alanine scanning suggested that LYS8, LYS73, and ASN43 contributions to interaction were not changed with minimisation; however ARG77 contributed more favourably in the minimised version. It has been suggested that this residue originally

contributes unfavourably because of a large desolvation penalty associated with burying this residue at the interface (Kajander et al., 2009). In the unminimised structure, SER76 is predicted to be the residue that contributes most favourably to the interaction by alanine scanning, however it was not predicted to participate in any interaction by the PIC. In the minimised structure, alanine scanning predicted that this residue contributed a lot less favourably to the interaction and paradoxically was also predicted by the PIC to participate in an ionic interaction.

Table 4.2: Comparison of the interacting residues in both the original and refined versions of 3UQ3 for the TPR2A region only (i.e ScHopTPR2A complexed to ScHsp-90 C-terminal peptide MEEVD).

Model	Method	Hydrophobic interactions	Main-side chain Interactions	Side-Side Chain Interactions	Ionic Interactions	Aromatic sulphur Interaction
Original	**PIC**	4	12	12	2	1
	Interacting residues found in AlaScan	A; 273TYR, ***285TYR***. B; MET706, VAL709	A; LYS266, ***ASN270***, TYR273, ASN300, ARG341, GLU307. B; GLU707, GLU708, ASP710	A; ***ARG276***, LYS337, ARG341, ASN344. B; ***MET706***, GLU707, ASP710	A; ***ARG276***, LYS337, B; ***MET706***, GLU707,	A; ***TYR273***. B; ***MET706***.
Minimised	**PIC**	4	10	11	6	0
	Interacting residues found in AlaScan	A; 273TYR. B; MET706, VAL709	A; LYS266, TYR273, ASN300, ARG341, GLU307. B; GLU707, GLU708, ASP710	A; LYS337, ARG341, ASN344. B; GLU707, ASP710	A; LYS337, ***ARG341***, ***LYS374***. B; GLU707, ***GLU708***, ***ASP710***	N/A

As can be seen from Table 4.2, it's clear to see that minimisation of 3UQ3 resulted in fewer interactions between TPR2A and the Hsp90 C-terminal peptide, and fewer still of the residues involved in these interactions were detected by alanine scanning to be important to complex formation. However, the minimised version did possess more ionic interactions than the unminimised version. The residues that contributed most significantly to interaction are the "carboxylate binding clamp" residues LYS266, LYS337 and ASN300, in both the original and minimised versions.

Most importantly, minimisation resulted in the loss of the aromatic sulfur interaction between TYR273 in HsHop and MET706 in Hsp90 that is thought to assist HsHop in distinguishing between HsHsp90 and HsHsc70 C-terminal peptide binding (Kajander et al., 2009). Alanine scanning indicates that minimisation causes both these amino acids to contribute less favourably to the interaction. This clearly reaffirmed that minimisation deteriorated this template complex, and would likely deteriorate the complex and quality of homology models built on this template.

Kajander et al (2009) noted that when the structures of TPR1 and TPR2A domains in complex with their C-terminal peptides are aligned, the Hsp70-peptide was in a conformation that could not fit in into a binding cavity in TPR2A, as it does into the TPR1 binding cavity. This specificitydisplayed by TPR2A is mainly as a result of the bulky methionine of Hsp90C-terminal MEEVD peptide better filling a cavity in the TPR2A convex surface than the corresponding isoleucine in the Hsp70 C-terminal GPTIEEVD peptide (Kajander et al., 2009; Schmid et al., 2012).

Table 4.3: Comparison of the interacting residues in both the original and refined versions of 3UQ3 for the TPR2B region only. (i.e ScHopTPR2B complexed to ScHsp-70 C-terminal peptide EVD).

Model	Method	Hydrophobic interactions	Main-Side Chain Interactions	Side-Side Chain Hydrophobic Interactions	Ionic Interactions
Original	**PIC**	3	7	0	2
	Interacting residues found in AlaScan	A; PHE407, TYR419. C; VAL709	A; ARG400, ASN435, ***ARG469***. C; ASP710	N/A	A; ***LYS404***, ARG469. C; ASP710
Minimised	**PIC**	3	5	3	2
	Interacting residues found in AlaScan	A; PHE407, TYR419. C; VAL709	A; ARG400, ASN435. C; ASP710	A; ***ARG465***. C; ***ASP710***	A; ***ARG465***, ARG469. C; ***GLU708***, ASP710

It is clear to see from Tables 4.3 and 4.4 (representing the ScHop TPR2B half of 3UQ3 and 3UPV, respectively) that there is some discrepancy between the number of interactions reported for both models, even though the interactions being represented are essentially the same (that of ScHop TPR2B in complex ScHsp70 C-terminal peptide), and were produced by the same authors, in the same manner, in the same species (Schmid et al., 2012). The most obvious reason for this is that the C-terminal fragment bound in 3UQ3 is EVD whereas the C-terminal fragment bound in 3UPV is PTIEEVD. There are two other possible reasons for these differences. The first is that it is possible that TPR2 may bind Hsp70 C-terminal peptide less strongly after binding Hsp90 C-terminal peptide owing to conformational changes in the S-shaped backbone. It is also likely that more interactions are detected in 3UPV as this model was of higher resolution (1.60 Å) than 3UQ3 (2.60 Å) and was a more accurate representation of the native structure.

Table 4.4: Comparison of the interacting residues in both the original and refined versions of 3UPV. (i.e ScHopTPR2B complexed to ScHsp-70 C-terminal peptide PTIEEVD).

Model	Method	Hydrophobic interactions	Main-Side Chain Interactions	Side-Side Chain Hydrophobic Interactions	Ionic Interactions
Original	**PIC**	3	11	9	4
	Interacting residues found in AlaScan	A; PHE273, TYR285. B; VAL655	A; ***ARG266***, ASN301, ARG335, GLU371. B; THR651, ASP656, GLU654	A; ***LYS308***, ***ARG331***, GLU371. B; THR651, ***GLU653***, ASP656	A; LYS308, ARG331, ARG335. B; GLU654, GLU653, ASP656
Minimised	**PIC**	4	10	5	4
	Interacting residues found in AlaScan	A; PHE273, TYR285. B; ***VAL652***, VAL655	A; ASN301, ARG335, GLU371. B; THR651, GLU654, ASP656	A; ***ARG335***, GLU371. B; THR651, ASP656	A; LYS308, ARG331, ARG335. B; GLU654, GLU653, ASP656

The minimisation of SchmidCYS model, depicting TPR2 (unbound to Hsp90 and Hsp70 C-terminal peptides) in complex with Hsp90 M and C domains, resulted in a large increase in interactions as well as the number of interacting residues involved in complex formation (see Table 4.5). There was only one Main-Main hydrophobic interaction loss. This is likely as a result of the method of producing this template (discussed in Chapter 3, Section 3.4.3). The docking of these two proteins with each other was rigid and relaxation of the side chain rotamers introduced more points of interaction.

Table 4.5: Comparison of the interacting residues in both the original and refined versions of SchmidCYS (i.e ScHopTPR2 complexed to ScHsp-90 M and C domains).

Model	Method	Hydro-phobic interaction	Main-Main chain interaction	Main-side chain Interaction	Side-Side Chain Interaction	Ionic Interaction	Cation-Pi interaction
Original	**PIC**	3	1	6	11	10	1
	Interacting residues found in AlaScan	A; TRP300 B; VAL416	A; THR462 B; THR368	A; ASN298, LYS449, B; LEU443, ARG376, SER445	A; CYS411, GLU446, ASP452, GLU453, GLU412, ASP302, B; LYS417, LYS383, LYS380, ARG376, LYS337, GLU379	A; GLU287, ASP302, GLU412, ASP452, GLU453, ASP459, GLU466, B; ***ASP370***, HIS366, ***LYS337***, ARG376, LYS380, LYS383, LYS417,	A; TRP300 B; ARG436
Mini-mised	**PIC**	4	0	18	16	14	1
	Interacting residues found in AlaScan	A; TRP300 B; VAL416, ***LEU440***	N/A	A; ASN298, LYS449, ***MET464***, ***TRP300***, ***ASN298***, B; LEU443, SER445, ***ARG367***, ARG376, ***LEU443***	A; CYS411, ***GLU415***, GLU412, GLU446, ***GLU287***, ASP302, ***GLU301,*** ***MET464,*** ***TRP300***, B; LYS417, ***LYS424***, LYS383, LYS380, ***LYS333***, ***GLU448***, ***ARG367***, ARG376, GLU379	A; GLU287, ***GLU301***, ASP302, GLU412, GLU453, ASP459, GLU466, ***ARG326***, B; GLU448, ***LYS333***, HIS366, ARG376, LYS380, LYS383, LYS417, ***LYS424***	A; TRP300 B; ARG436

4.3.2: Protein-Protein Surface Interactions and Characterisation

Table 4.6: Comparison of the interacting residues in the refined PfHopTPR1 Model (36) and its human homolog 1ELW.

Model	Method	Hydrophobic interactions	Main-Side Chain Interactions	Side-Side Chain Hydrophobic Interactions	Ionic Interactions
1ELW_mod	**PIC**	8	13	2	5
	Interacting residues found in AlaScan	A, ***LEU15***, ***TYR27***, ***PHE84***. C; VAL8, ***ILE6***	A; ***LYS8***, ASN12, ASN43, ***LYS73***, ARG77. C; ASP12, GLU10, GLU9	A; ***LYS73***. C; ASP12	A; ***LYS73, ARG77***. C; ASP12, GLU10, GLU9
Minimised PfHopTPR1. B99990036.pdb	**PIC**	6	13	0	2
	Interacting residues found in AlaScan	A; ***PHE18***, B; ***PRO123***, VAL125, VAL128	A; ASN15, ASN46, ARG80 B; GLU126, GLU127, ASP129	N/A	B; ASP129
Conserved	**PIC**	**3**	**9**	**N/A**	**1**
Aligned	**PIC**	**3***	**9****	**N/A**	**1**

*However, there is also a displaced PHE57 (undetected by alanine scanning) in *P.falciparum*, in the same region as PHE84 in 1ELW (but on adjacent helix) performing the same function.

** No LYS interactions in *P. falciparum* model confirmed by AlaScan

From Table 4.6, it is evident that there are several differences between homologous *P. falciparum* and *H. sapiens* TPR1:Hsp70 C-terminal complexes. The active sites of both proteins have several aligned interactions, i.e. those interactions of the same type occurring between different residues (or same residues but different atoms) that align in both species' structures. There are slightly fewer conserved interactions, i.e. those interactions of the same type occurring between identical residues (and atoms) that align on both species' structures.

As is discussed by Scheufler, et al (2000) most of the direct hydrogen bonding interactions from the TPR1 region are to the Hsp70 C-terminal peptide backbone. PIC results show that the same is true for the the PfHop model. It was previously concluded that this meant that TPR1:Hsp70 C-teminal complex formation did not rely on sequence-specific features of the Hsp70 peptide in the human complex (Scheufler et al., 2000). However, *in-silico* alanine scanning of the complex indicated that mutation of any of the Hsp70 peptide residues in both complexes indicated that all the residues in –IEEVD in HsHsp70 and –VEEVD in PfHsp70 result in complex destabilisation for both species when mutated to alanine (Table 4.6).

In the human complex, ARG77 of TPR1 plays a key role in binding the backbone of the peptide. Its guanidinium group makes three direct hydrogen bonds with the carbonyls of GLU9 and VAL10 in the EVD C-terminal (Scheufler et al., 2000). In PfHop, in the aligned position, ARG80 is only making two direct hydrogen bonds with the residue homologous to HsHop GLU9, which is to the carbonyl of GLU127. An additional hydrogen bond to the carbonyl of GLU9 is mediated by a tightly bound water molecule positioned by LYS50 in TPR1 in human (Scheufler et al., 2000). Neither the PIC nor AlaScan programs detected this LYS50-GLU9 interaction in 1ELW; however, it was detected by these programs in the refined version of this structure (see Table 4.1). Owing to the fact that all waters are removed from the template prior to homology modelling and that the homologous residue in *P.falciparum* is SER53, it is not surprising that an aligned interaction was not predicted by the PIC. However, alanine scanning results indicated that mutation of the SER53 to alanine may result in complex disruption, indicating that the residue in this position is important. This highlights the necessity of using more than one method to analyse individual residue interactions in complex models.

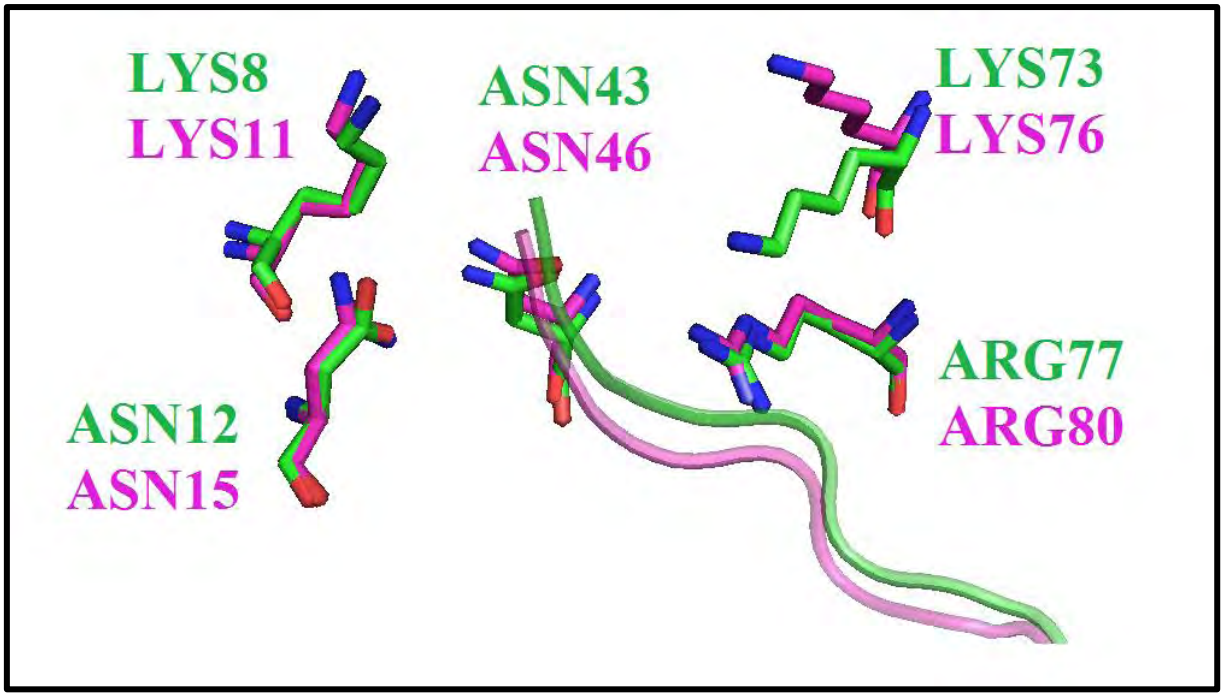


Figure 4.2: 1ELW (green) aligned to minimised PfTPR1 model 36 (pink) and visualised in PyMOL, showing the "carboxylate clamp" residues (stick representation) interacting with the C-terminal -EEVD of Hsp70/Hsc70 (cartoon representation).

Importantly, the highly conserved "carboxylate clamp" (Scheufler et al., 2000), is present in both the human and malarial complex (Figure 4.2). In the *P.falciparum* model, the PIC predicts ionic protein interactions for LYS11 and LYS76 to ASP129 however; alanine scanning does not predict that these lysines are critical to complex stabilisation hence their exclusion form the Table 4.6. For the homologous interaction in humans, LYS8 and LYS73 to Hsc70's ASP12, alanine scanning predicts that the mutation of these lysines to alanine will destabilise the complex.

Table 4.7: Comparison of the interacting residues in the PfHopTPR2A model 71 and the homologous HsHopTPR2a model 13.

Model	Method	Hydrophobic interactions	Main-side chain Interactions	Side-Side Chain Interactions	Ionic Interactions	Aromatic sulphur Interaction
HsTPR2A13	**PIC**	4	14	9	5	1
	Interacting residues found in AlaScan	A; TYR12, TYR24, B; MET264, VAL267	A; ***LYS5***, ASN9, TYR12, ASN40, GLU47, LYS77, ARG81, B; GLU265, GLU266, ASP268	A; LYS15, ***THR36***, LYS77, ARG81, ASN84, B; MET264, GLU265, ASP268,	A; ***LYS5***, LYS77, ARG81, B; GLU265, ***GLU266***, ASP268	A; TYR12, B; MET264
PfTPR2A14	**PIC**	4	10	3	2	1
	Interacting residues found in AlaScan	A; TYR12, TYR24. B; MET245, VAL248	A; ASN9, TYR12, ASN40, GLU47, ARG81, B; GLU246, GLU247, ASP249	A; LYS15, LYS77, ARG81, B; MET245, GLU246, ASP249	A; LYS77, ARG81 B; ASP249, GLU246,	A; TYR12. B; MET245.
Conserved	**PIC**	**4**	**10**	**3**	**2**	**1**
Aligned	**PIC**	**4**	**10**	**3**	**2**	**1**

From Table 4.7, it is evident that there are very small differences between homologous *P. falciparum* and *H. sapiens* TPR2A:Hsp90 C-terminal complexes. Generally, the human complex has more sites of interaction and all the active sites of *P. falciparum* are identical to their aligned human interactions. It is also clear to see from Figure 4.3, that all the carboxylate binding clamp residues in both species are in very similar orientations for complex formation.

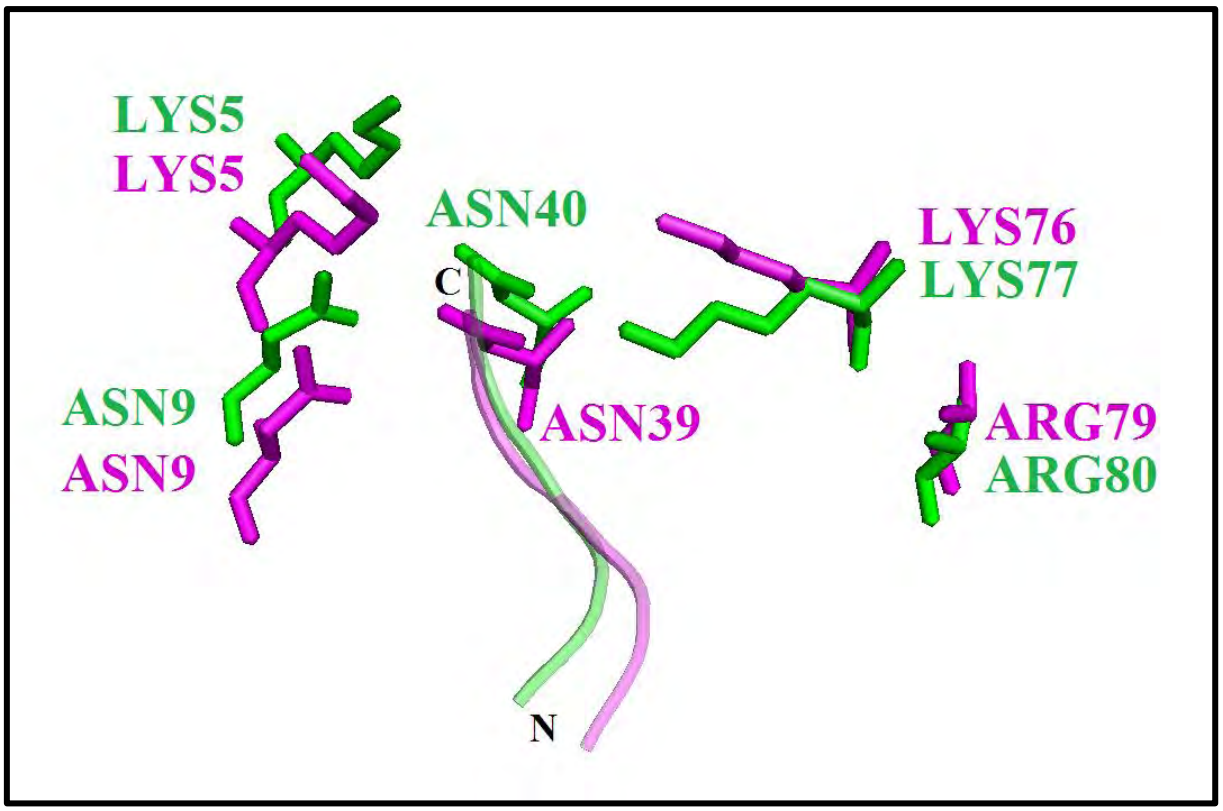


Figure 4.3: Comparing interaction sites in HsTPR2B model 13 (green) aligned to PfTPR2B model 14 (pink) and visualised in PyMOL. The "carboxylate binding-clamp" residues (stick representation) are interacting with the C-terminal -MEEVD of Hsp90 (cartoon representation).

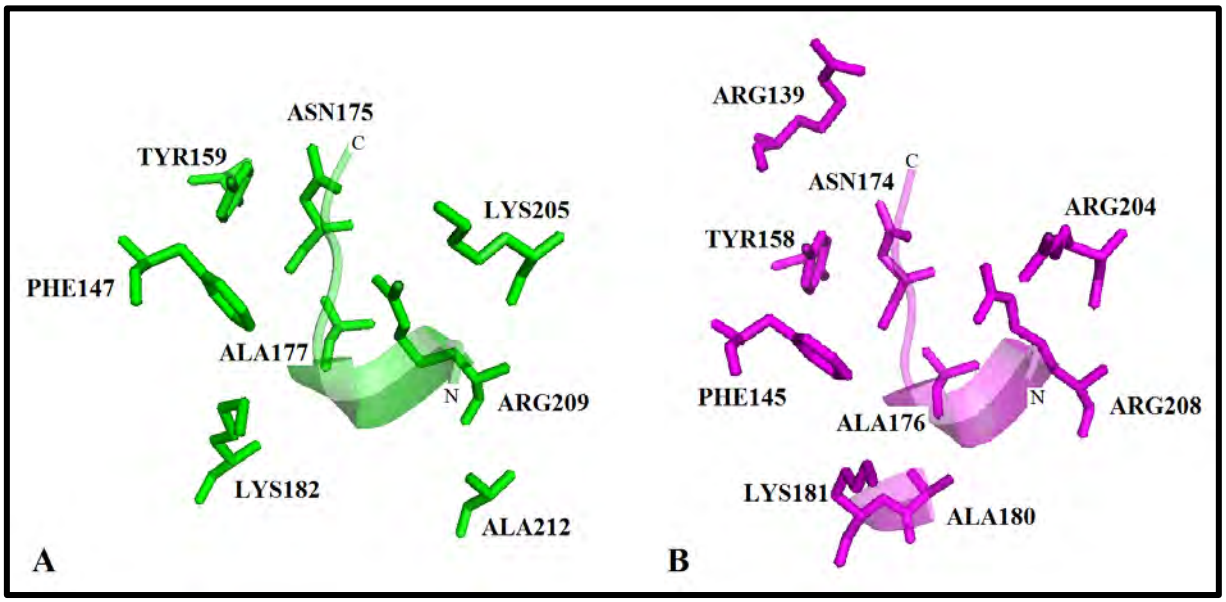


Figure 4.4: Comparing interaction sites in HsTPR2B model 13 (green) aligned to PfTPR2B model 14 (pink) and visualised in PyMOL. The interacting residues (stick representation) interacting with the C-terminal -PTIEEVD of HsHsc70 and the C-terminal -PTVEEVD of PfHsp70, respectively (cartoon representation).

Table 4.8: PIC results comparing binding residues in human and malarial TPR2B:Hsp70-PTVEEVD complexes.

Model	Method	Hydrophobic interactions	Main-side chain Interactions	Side-Side Chain Interactions	Ionic Interactions	Aromatic sulphur Interaction
HsTPR2B13	**PIC**	4	5	1	4	0
	Interacting residues found in AlaScan	A; PHE147, TYR159, B; ***ILE271***, VAL274	A; ARG209, ASN175 C; GLU273, THR270, ASP275	A; LYS182, C; GLU272,	A; LYS182, LYS 205, ARG209 C; GLU272, GLU273 ASP275	N/A
PfTPR2B14	**PIC**	3	8	1	4	0
	Interacting residues found in AlaScan	A; PHE147, TYR159, ***ALA178**** C; VAL255	A; ***LYS140***, ASN175, ARG209. C; THR251, GLU254, ASP256	A; LYS182, C; GLU253,	A; LYS182, ARG 205, ARG209 C; GLU253, GLU254, ASP256	N/A
Conserved	**PIC**	**3**	**1**	**1**	**4**	**N/A**
Aligned	**PIC**	**3**	**5****	**1**	**4**	**N/A**

*ALA178 in *P.falciparum* detected by alanine scanning, but human homolog ALA178 is not.

**There are four bonds in *P. falciparum* from NH1 of ARG209 to O of THR251 and GLU254, but for human, there is an analogous interaction occurring from NH2 of ARG209 to O of THR270 and GLU273.

From Table 4.8, it is evident that there are very small differences between homologous *P. falciparum* and *H. sapiens* TPR2B:Hsp70 C-terminal complexes. As shown in Figure 4.4, the assembly of interacting residues are very similar, with the exception of the variability at the LYS205 and ARG204 position. Generally, the complexes have an equal amount of interactions and again, almost all the active site interactions of *P. falciparum* are identical to their aligned human interactions. In the few cases where interactions are not identical, there are aligned interactions occurring between identical residues, but different side chain nitrogen groups become the H-bond donor.

Table 4.9: PIC results comparing binding residues in human and malarial Hsp90-M&C-domains:HopTPR2 complexes.

Model	Method	Hydrophobic interactions	Main-side chain Interactions	Side-Side Chain Interactions	Ionic Interactions	Cation-Pi interactions
Minimised HsComplex model 7	**PIC**	3	3	11	6	1
	Interacting residues found in AlaScan	A; TRP26 B; MET554, TYR578, LEU586	A; LYS20, ASN24, TRP26, B; THR558, LEU581	A; GLU13, GLU138, ASP185, ARG189 B; GLN514, ARG525, LYS555	A; GLU13, ***ASP28***, LYS137, GLU138, LYS141, GLU179, B; LYS518, ARG525, ***LYS555***	A; TRP26 B; ARG574
Minimised Pfmulti model 102	**PIC**	4	5	8	7	0
	Interacting residues found in AlaScan	TYR26 B; TYR569, ALA589, LEU590, LEU593	A; LYS26, ALA25, TYR26, ASN192 B; ARG519, LYS566, ASP570, GLU595	A; ASN13, ASP141, ASP181, LYS182 B; ARG519, LYS526, ARG530, GLU533, LYS567	A; ***ASP28***, ASP141, ASP181, GLU191, LYS182, ASP226 B; ARG519, ARG522, LYS526, ARG530, LYS526, ARG530, GLU533, ***LYS567***	N/A
Conserved	**PIC**	**0**	**0**	**0**	**1**	**0**
Aligned	**PIC**	**1**	**1**	**2**	**3**	**0**
Region of Occurrence		**Only TPR2B and Hsp90M**	**Primarily TPR2B and Hsp90M**	**Primarily Linker, TPR2A and Hsp90C**	**Primarily Linker, TPR2A and Hsp90C**	**Only Linker and Hsp90C**

The convex binding sites on HsHop TPR2 in complex with the HsHsp90 M and C domains share almost no conservation of interactions (in terms of interacting residues) to their

counterparts in PfHop TPR2 in complex with PfHsp90 M and C domains and a relatively small percentage of those interactions were even aligned (see Table 4.9). The only interaction that was conserved between the two species was an ionic interaction (bold italics, Table 4.9) between ASP28 (in both species Hsp90 C domain) and a lysine (LYS555 and LYS567 in HsHop TPR2 and PfHop TPR2, respectively). Additionally, alanine scanning identified more residues integral to complex formation in *P. falciparum* than in human Hop.

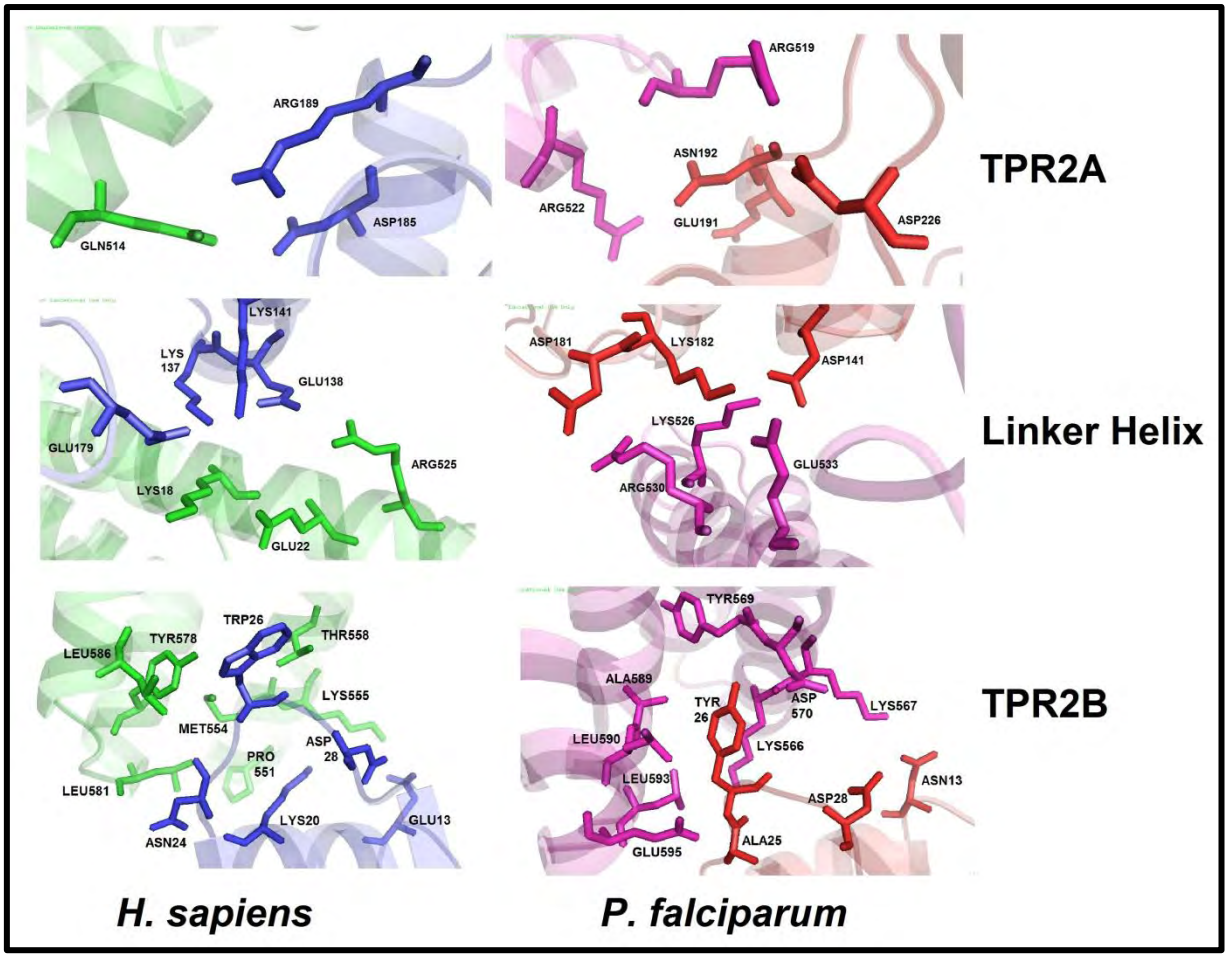


Figure 4.5: HsTPR2A model 13 (green) and PfTPR2A model 14 (pink) and visualised in PyMOL, showing the interacting residues (stick representation) interacting with the Hsp90 C and M domains (blue and red for human and *P. falciparum*, respectively).

The regions in which interaction was observed as well as the numbers of each interaction type, however, seemed to be well conserved (see Table 4.9 and Figure 4.5). Overall it was plain to see that this interaction is important to complex formation in both species; however the interacting residues showed far more variability than the interacting residues on the concave surfaces of TPR1 and TPR2A&B.

4.4 Binding Energies of Several Complexes

4.4.1 TPR1 Complexes

Table 4.10 shows that the comparative interaction and binding energies in TPR1 complexes are higher in the PfHop model, indicating the PfTPR1 domain has lower affinity for the C-terminal motif of its Hsp70 partner than does the HsTPR1 domain for its respective Hsc70 partner. With regards to cross-species binding of HsHop to Hsp70-x GPTVEEVN C-terminal motif, it is ambiguous as to which species has the higher affinity for Hsp70-x, as the interaction energy for the HsHop complex is higher, while the overall binding energy is lower.

Table 4.10: Binding energy summary for single template complex models.

Template		**Rosetta Energies**				
		H. sapiens			***P. falciparum***	
1ELW Chain A and C Human		**One Complex - TPR1:GPTIEEVD**			**One Complex - TPR1:GPTVEEVD**	
XRD Scheufler *et al*., 2000 TPR1 Motif	- 1.60 - 0.180 - 0.215	Template: 1ELW_mod			PfHopTPR1 model 36	
		Interaction Energy:	-10.709		Interaction Energy:	-4.515
		Binding Energy:	42.112		Binding Energy:	299.633
		-			-	
1ELW Chain A and C Human		**One Complex - TPR1:GPTVEEVN**			**One Complex - TPR1:GPTVEEVN**	
XRD Scheufler *et al*., 2000 TPR1 Motif	- 1.60 - 0.180 - 0.215	HsHopTPR1N model 1			PfHopTPR1N model 50	
		Interaction Energy:	7.791		Interaction Energy:	-4.91
		Binding Energy:	208.804		Binding Energy:	349.301
M&C Domain Hsp90 & HopTPR2 Yeast		**One complex**			**None**	
Docked Spin-Labelled complex Schmid *et al*., 2012 TPR2AB (Convex)		Hs_complex_single model 7			N/A	
		Interaction Energy:	-21.402			
		Binding Energy:	-195.339			

What is also important to note is that the overall affinity of HsHopTPR1 for its Hsc70-1 C-terminal motif is much greater than for the PfHsp70-x motif, so this motif may be out competed for binding in the red blood cell cytoplasm. Interestingly, overall affinity of PfHopTPR1 for its Hsp70-1 C-terminal motif is greater than for the PfHsp70-x motif, so this motif may be outcompeted for binding within the parasite cytoplasm.

4.4.2 TPR2 Complexes

The comparative interaction and binding energies in TPR2 models are higher in both *P. falciparum* HopTPR2:Hsp70-1:Hsp90 and HopTPR2:Hsp70-x:Hsp90 complexes (see Table 4.11). This indicates the PfTPR2A&B C-terminal binding domains have lower affinity for both C-terminal motifs of its Hsp70 and Hsp90 partners than does the HsTPR2 domain for its respective Hsc70 and Hsp90 partners. In all models, the HopTPR2A has stronger affinity for its Hsp90 C-terminal motif partner than does the TPR2B domain for its Hsp70-1 partner. This is consistent with both recent and older findings, both *in vitro* (Scheufler et al., 2000; Schmid et al., 2012) and *insilico* (Kajander et al., 2009).

What is interesting to note is that in HsHop, all TPR regions complexed to their respective motifs have lower binding energies than their *P. falciparum* counterparts. However, the interaction and binding energy are lower for the *P. falciparum* complex between HopTPR2 and Hsp90 M and C domains (Table 4.11), than for the human counterpart (Table 4.10).

Table 4.11: Binding energy summary for multi template complex models.

Templates		**Homology Models – Best DOPE-Z Score of 100**					
A	**B**	***H. sapiens***				***P. falciparum***	
M&C Domain Hsp90 & HopTPR2 And 4GCO *C. elegans*		**None**				**One complex**	
Spin-labelled Scheufler *et al*., 2000 TPR2	Chain A TPR2B	N/A				Pfmulti model 102	
						Interaction Energy:	-24.59
						Binding Energy:	-282.631
3UQ3 Chain A, B and C 3UPV Chain A and B Yeast		**Two complexes** - **TPR2A:MEEVD** - **TPR2B:PTIEEVD**				**Two complexes** - **TPR2A:MEEVD** - **TPR2B:PTVEEVD**	
XRD Schmid *et al*., 2012 TPR2AB Motifs	- 2.60 - 0.222 - 0.279	HsTPR2ab2yeast model 13				PfTPR2ab2yeast model 61	
			2A	**2B**			**2A**	**2B**
		Interaction Energy:	-7.746	8.996		Interaction Energy:	-6.533	41.979
		Binding Energy:	-3640.731	-3620.463		Binding Energy:	628.2	653.952
		Overall Interaction Energy:	1.25			Overall Interaction Energy:	35.446	
		Overall Binding Energy:	-3624.447			Overall Binding Energy:	652.644	
3UQ3 Chain A, B and C 3UPV Chain A and B Yeast		**Two complexes** - **TPR2A:MEEVD** - **TPR2B:PTVEEVN**				**Two complexes** - **TPR2A:MEEVD** - **TPR2B:PTVEEVN**	
XRD Schmid *et al*., 2012 TPR2AB Motifs	- 2.60 - 0.222 - 0.279	Top Model: HsTPR2ab2N model 83				Top Model: PfTPR2ab2N model14	
			2A	**2B**			**2A**	**2B**
		Interaction Energy:	-9.813	25.603		Interaction Energy:	8.717	93.106
		Binding Energy:	45.657	81.834		Binding Energy:	627.293	707.019
		Overall Interaction Energy:	15.79			Overall Interaction Energy:	101.822	
		Overall Binding Energy:	78.653			Overall Binding Energy:	721.523	

4.5 Correlating Sequential and Structural Features with Interaction Sites

Analysis of most of the MEME motifs, particularly those describing fungal and protozoan sequences (discussed in Chapter 2, see also Appendix 1), indicated a periodicity of conserved residues at approximately every third or fourth site. As this is roughly the periodicity of an alpha-helix turn, it was suspected that this periodicity of conservation reflected conserved surfaces in the Hop structure. To investigate this, conserved residue sites identified in fungal and protozoan motifs were mapped to the corresponding sites on the PfHop structures discussed in Chapters 3 and 4.

Using motifs 2, 9 and 10 from the twenty motif search (i.e. that discussed in Chapter 2), the residues with bit scores higher than 3 were mapped to the minimised PfTPR1 model 36, and displayed as a surface. Interestingly the concave surface is very well conserved and is almost completely mapped with these well conserved (bit score > 3) regions (see Figure 4.6 A).

The convex surface of PfTPR1 is less well conserved, however there are two blocks of well conserved sites (block B1 & B2, Figure 4.6 C2) that map two "patches" of the convex surface (region B1 and B2, Figure 4.6 B). TPR1 is thought to interact with other proteins by means other than the concave binding site residues in the human model (Scheuflur et al, 2000; Kajander et al, 2009), and may also work in concert with the DP2 domain to stabilise the interaction between Hsp70 bound to client when in complex in the yeast model (Schmid et al, 2012). It is possible that these "patches" of conserved residues on the convex side of TPR1 could identify these as yet uncharacterised regions of interaction in Hop TPR1. Only further experimental studies may confirm this.

To the author's knowledge, this is the first reported instance of using MEME results to identify conserved surface features and relate them to structures in this way. This may turn out to be a useful method for examining the suspected conserved and functional features of protein structures in the future.

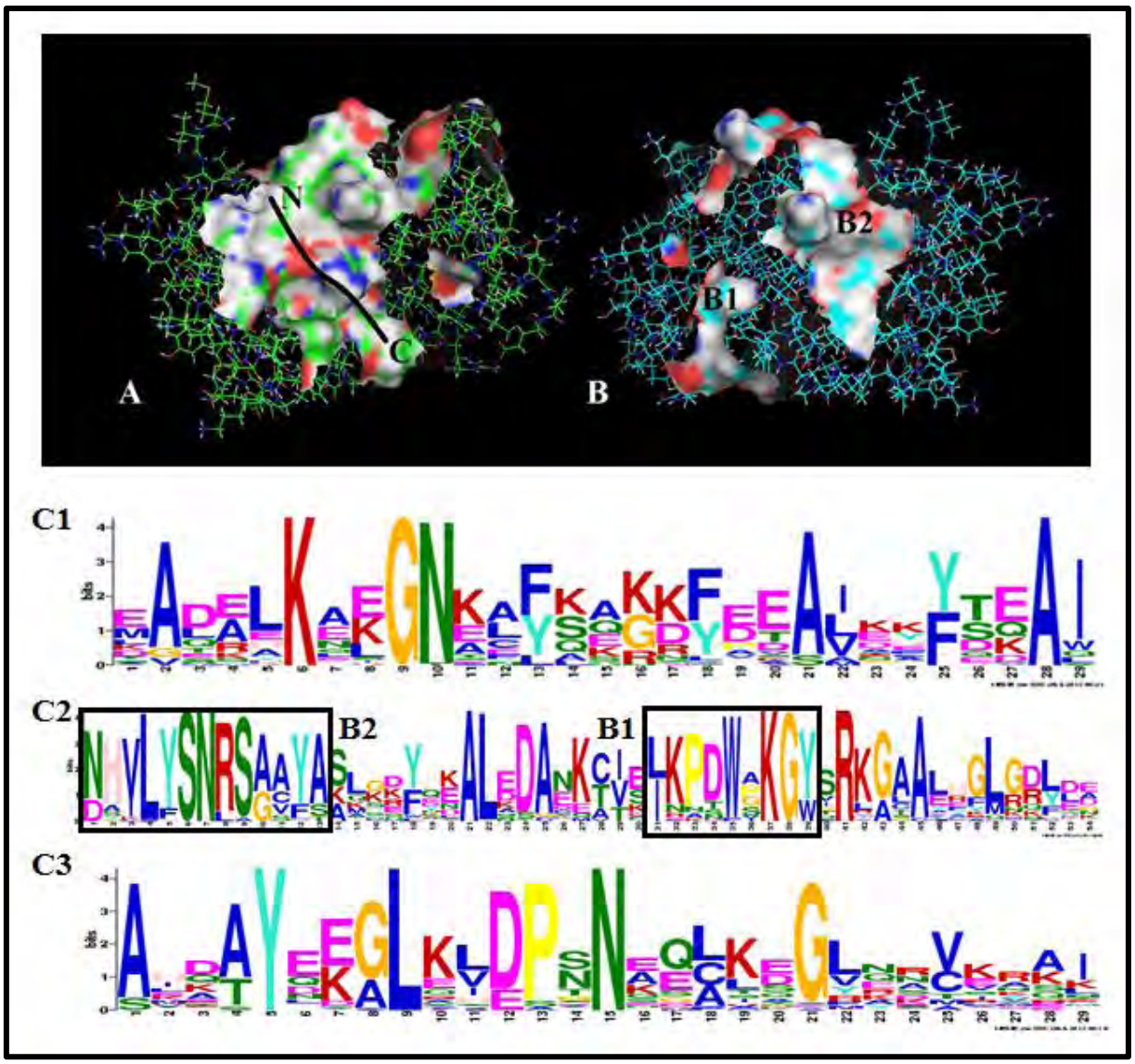


Figure 4.6: Residues with a bit score > 3 in Minimised PfTPR1 model 36 (chain A) selected and displayed as surface in PyMOL. A) The concave surface of the TPR1 domain (green alpha-carbon wireframe model) with the cartoon representation (black) of PfHsp70-1 C-terminal motif. B) The convex surface (cyan alpha-carbon wireframe model). B1 and B2 correspond to 2 stretch of well conserved residues and are highlighted in C2. C1) Motif 10 from Figure 2.6. C2) Motif 2 from Figure 2.6. C3) Motif 9 from Figure 2.6.

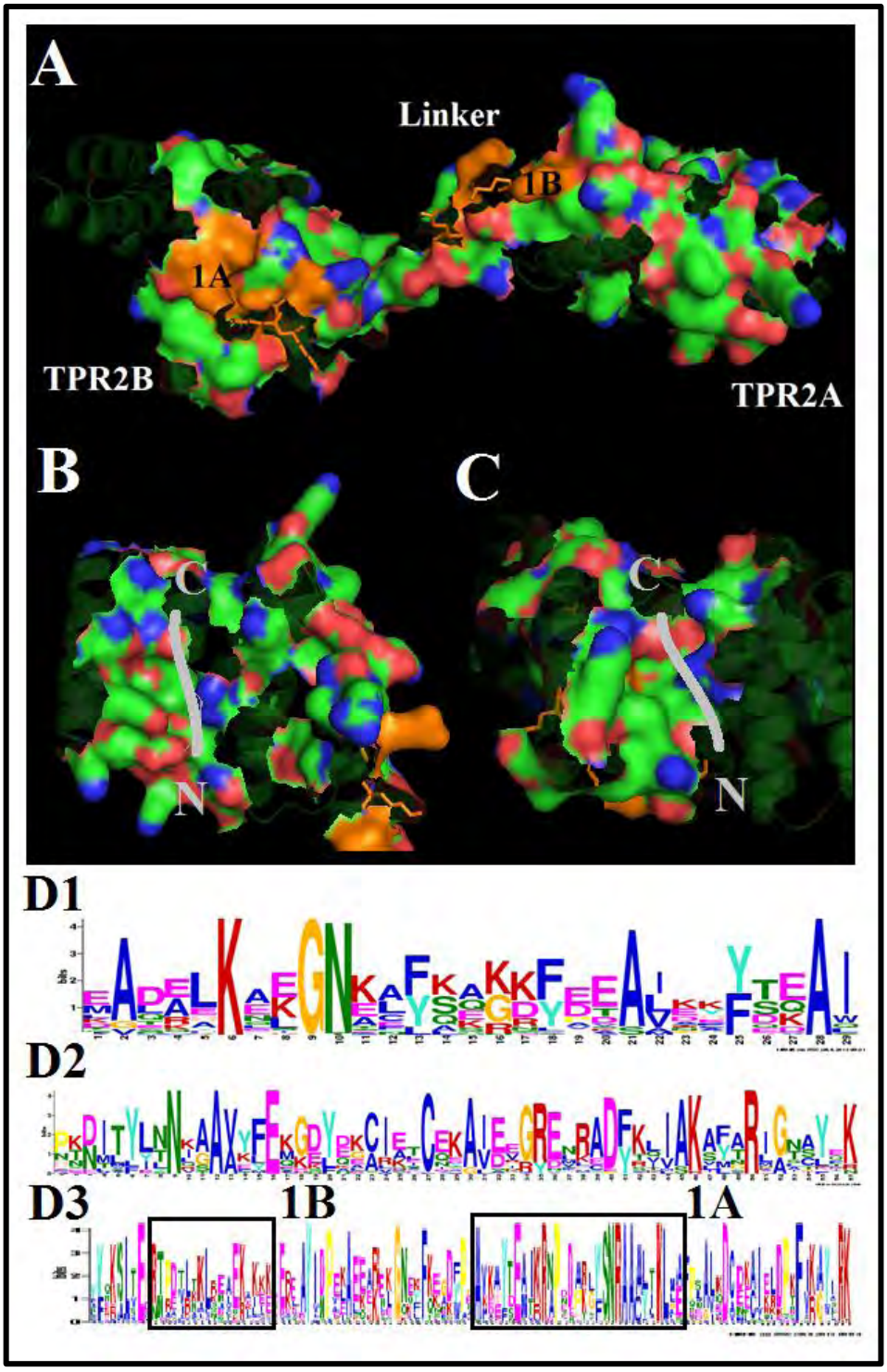


Figure 4.7: Residues with a bit score > 2.5 in Pfmulti model 102 (chain B) selected and displayed as surface in PyMOL. A) The convex surface of the TPR2A&B domain (green alpha-carbon ribbon) with the stick and surface representation (orange) of Hsp90 binding residues. 1A and 1B correspond to 2 stretches of well conserved residues and are highlighted in D3. B) The concave surface of TPR2A. B) The concave surface of TPR2B. D1) Motif 10 from Figure 2.6, representing the N-terminal region of TPR2A. D2) Motif 4 from Figure 2.6, representing the middle of TPR2A. D3) Motif 1 from Figure 2.6, representing the C-terminal of TPR2A, linker helix and TPR2B regions.

Using motifs 1, 4 and 10 from the twenty motif search, the residue sites with a bit score higher than 2.5 were mapped to the Hop half of Pfmulti model 102 (chain B), and displayed as surface (Figure 4.7). Not surprisingly, the concave surfaces of TPR2A and for TPR2B were well conserved and were almost completely mapped with these well conserved (bit score > 2.5) regions. At a bit score of greater than 3, only four of the residues in PfHop that bind PfHsp90 M and C Domains in Pfmulti model 102 (ARG530, TYR569, ALA589 and LEU593) are represented. These are all residues identified by alanine scanning to be important to complex formation (Table 4.9). At a bit score of greater than 2.5, a further four residues (ARG522, ASP570, LEU590 and GLU595) involved in the interaction are represented. Again these are all residues identified by alanine scanning to be important for complex formation (Table 4.9).

4.6 Conclusions

The primary conclusions of this chapter are two-fold; firstly, as discussed in Chapter 3, minimisation can severely impact the structure of models and secondly, that the concave interaction sites in the TPR domains are marginally better conserved than the convex interacting sites. From Section 4.3, it is observed that when a model (or template) is of low quality it appears that minimisation introduces new points of interaction and alanine scanning detects greater numbers residues that are important to the interaction. However, when the model (or template) is of sufficient quality not to warrant minimisation, there is either a small loss or gain in interactions. If a model is of good quality, it is probably advantageous to leave it alone, as minimisation will result in the risk of introducing bias toward a certain conformation (as discussed in Chapter 3, Section 3.1.4). If a model is of low quality, it may well be with the risk of introducing bias in order to identify a wider range of interaction sites. Similar refinement advantages and disadvantages have been discussed elsewhere in the literature (Tastan Bishop & Kroon, 2012; Vyas et al., 2012).

Overall, from Section 4.4, it is clear that the concave binding site interactions and residues (primarily the “carboxylate binding clamp” residues) are exceptionally well conserved between the two human and *P. falciparum*Hop sequences. In fact, Section 4.6 suggests that the entire concave surface of the TPR domain in general is extremely well conserved. In contrast, Section 4.4 also shows that the convex surfaces of the TPR2 domains involved in complex formation with the Hsp90 M and C domains display much more variability in

residue composition between the two species. However, the sites of interaction remain the same. This would be the more plausible protein-protein interaction site to target with selective drugs in order to combat malaria.

Section 4.5 dealt with the question of whether there is potential for a cross-species protein complex to form, as discussed in Chapter 1, Sections 1.2 and 1.9. However it also allowed generalisations to be made about differences in protein affinities between each species. Overall, it appears that the human complexes have lower interaction and binding energies than do the *P. falciparum* complexes, with the exception of the TPR2:Hsp90 M and C domain complex, where the interaction and binding energies are slightly lower in *P. falciparum* than in human complexes. This is reflected in Table 4.9, which shows that this complex has a higher number of interactions (and residues favourably involved in interaction) in *P. falciparum* than in human.

It has long been known that *P. falciparum* can annex and manipulate the host cell's proteins for growth (Charpian et al, 2009). With regards to cross-species protein complex formation to occur, it is possible that PfHsp70-x may interact with human Hop. The results from Section 4.5.2 seem to support the likelihood of this interaction. The interaction and binding energies in the PfHopTPR2B:PfHsp70-x models are higher in *P. falciparum* than in the human HsHopTPR2B:PfHsp70-x complex, indicating that HsHop has a higher affinity for PfHsp70-x C-terminal peptide than PfHop, allowing PfHop to be outcompeted for binding in the host cell cytoplasm. Additionally, overall affinity of PfHopTPR1 for its Hsp70-1 C-terminal motif is greater than for the PfHsp70-x motif, so this motif may also be outcompeted for binding within its native parasite cytoplasm.

Chapter 5: Summary and Implications of Research

5.1 Project Results in Brief

It was clear from Chapter 2 that phylogeny-wide MSA shows good conservation of the Hop protein across all sequences analysed. Phylogenetic analysis at both the protein and gene level, while disagreeing on the order of outgroups, indicates that there is distinct grouping of Hop sequences into several units; vertebrate and invertebrate, fungal and protozoan. Overall, the most well conserved region of Hop is the DP2 region, indicating that this domain may have a more important functional role than previously thought. Future research efforts will need to focus on the experimental characterisation of this region.

On the other hand, the DP1 domain and linker region connecting DP1 and TPR2A domain appears to be the least well conserved region in the protein. MSA analysis of this area in the protein suggests this is the region of most insertions and deletions. Because the DP1 domain possesses no recognisable "DP" repeats and the long linker contains no proline repeats in the apicomplexan taxa (particularly in the *Plasmodium* genus), it is probable that this region will form a structurally and functionally distinct domain that is unique to this taxa (in comparison to all other taxa) and has yet to be characterised experimentally in PfHop. The production of poor quality models of PfDP1 discussed in Chapter 3 only reinforces this conclusion.

With regards to the TPR domains in all species, and apart from the lysine/arginine variability in the "carboxylate binding clamp residues" of TPR2B, the binding site residues for the concave surfaces (in fact, the whole concave surfaces) of the 3 TPR motifs domains are highly conserved across all taxa. The production of good to excellent quality models of human and *P.falciparum* TPR domains (in complex with their respective C-terminal peptide partners) discussed in Chapter 3 again reinforce this conclusion. In contrast, the convex surface residues of TPR2A&B that are involved in the interactions that allow Hop to bind Hsp90 in on surface regions of the M and C domains are less conserved between HsHop and PfHop. However, the sites of these interacting residues appear to be well conserved. Any further research that aims to find drugs that target Hop protein-protein interactions would likely have the most success targeting this interaction between PfHop and PfHsp90.

Chapter 4 suggests that HsHop TPR domains bind their C-terminal peptides more strongly than do PfHop TPR domains. This could be advantageous for chemotherapeutic drugs that target these concave surface TPR interactions, as it is possible that HsHop TPR interactions

will be less easily disrupted than their *P. falciparum* counterparts. While this is an interesting and exciting result, there were several limitations on the way these results were obtained (as discussed in Section 4.2). Further *in-silico* and experimental research will need to focus on simulating the physiological conditions in which these interactions take place in the uninfected human host cell, the parasite cell and the infected cell in the trophozoite stage.

Additionally, Chapter 4 provides evidence for the possibility of cross-species protein-protein interactions. While both HsHop and PfHop have lower affinities for the alternative PfHsp70-x C-terminal motif discussed in Chapters 1 and 4, HsHop has a greater affinity for the peptide than does PfHop. Again the validity of these results will need to be explored under simulated or experimental physiologically appropriate conditions. If it is found that such interactions do exist between *P. falciparum* and human Hsp and chaperone proteins within the infected host, this information would lead to exciting new protein interactions and biochemical pathways for drug development, so this idea definitely needs to be explored further.

5.2 Novel Approaches Used in this Project

According to the the author's current knowledge, the methods of comparing MEME/MAST block diagrams according to the phylogeny of the sequences used (Chapter 2), as well as using MEME sequence logos to identify conserved surface features and map them to structures (Chapter 4) are novel ways of using the MEME software. This may turn out to be a useful method for examining the suspected conserved and functional features of protein structures as well as the validation of protein trees (or even gene trees) in the future.

5.3 Hop as a Potential Drug Target

Overall this project addressed its aims of determining whether the PfHop protein and its role in other protein interactions would make a suitable drug target. The simple answer is yes. In reality, there is a lot that still needs to be understood about *P. falciparum*'s mechanism of growth within the infected host's red blood cell. A good summary of the most up to date research on trafficking of malarial proteins was recently published by Deponte et al (2012). However, this report mentions Hop only briefly and otherwise there is very little information available on the role of PfHop within the parasite as well as HsHop within the mature erythrocyte environment.

5.4 Project Expansion

There is much scope for the expansion of this project. While Hop in general has only been studied intensively for a relatively short period of time (approximately 10 years), it has been implicated in a very large number of pathways of interest to research in several high-impact diseases. These are thoroughly discussed in Chapter 1. There is no official collective publication on how vast the network of this protein as yet. Hop is clearly a small protein, but part of a big interaction network which is currently of a lot of interest to several aspects of disease research. Most high profile proteins or pathways are researched as such. However, each research group will tend to focus on their own separate pathways of interactions within their field of interest. While this is standard practise, it might be wise to change this paradigm of research to one that is at least peripherally aware of all the other interactions a particular group's protein of interest is involved in, both in other species and other cell types in the same species.

One method of reaching this holistic view of all interactions pertaining to a protein or class of proteins is already currently being developed. A bioinformatics PhD student, G. Salazar, based in Cape Town, is currently working on a user-friendly interface that visualises protein-protein interactions in a networked "bubble" viewer. The current prototype, which has been built using information on protein-protein interactions between multiple *Mycobacterium tuberculosis* proteins of interest to TB research and can be viewed here: http://biosual.cbio.uct.ac.za/biosual/tests/pinv/pinv.html. The author of this tool also documents the development of work in a blog, available here: http://biosual.blogspot.com/2013/01/biosual-viewer-first-prototype.html. Personal communication with this author at a conference earlier this year revealed that future aimsof the project involved outside parties being able to upload information into the network tool, and create their own networks.

It is for this reason that there the next step in the current work is a collaboration to combine this tool with a repository of information on Hop (and later, all Hsp proteins) protein-protein interactions, in various species as well as various organ and cell types. It would be available online as a user-friendly tool available to all those seeking information on current Hop research. In order to build on this experimentally, the role of Hop within *P. falciparum* should be more thoroughly characterised and the existence of HsHop within the mature red blood cell needs to confirmed, in order to design experiments that build on the current work's results.

Ideally, there needs to be attempts to create models of the full length structure of Hop in human yeast and *P. falciparum*, both in solution-state and in complexed conformations.

It may be possible to use the BUNCH assembly (computational assembly of several experimentally produced domain structures, and the linker region sequences) approach used by Romano et al (2009) to produce an accurate, full length working structure of HsHop. Analyses in Chapter 2 showed that (between XRD models of TPR1 and TPR2A, the probable polyproline II helix in the linker between mammalian DP1 and TPR2A, Cryo-EM models of TPR2A&B, and the homology models presented in this project) there are approximately only 10 residues in the linker between TPR2B and DP2 for HsHop that have not been mapped to a structure. Unfortunately, this approach would not be as feasible with PfHop. This is because the DP1 and long linker region have not been structurally or functionally characterised yet and its role in the formation of the Hsp chaperone complex as not been nearly as well described or investigated as that for yeast and human.

References

Abbas-Terki, T., Donzé, O., Briand, P.-A., & Picard, D. (2001). Hsp104 interacts with Hsp90 cochaperones in respiring yeast. *Molecular and Cellular Biology*, *21*, 7569–7575.

Accelrys Software Inc., Discovery Studio Modeling Environment, Release 3.5, San Diego: Accelrys Software Inc., 2007.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell*. (S. Gibbs, Ed.) (4th ed., p. 1463). New York: Taylor & Francis Group.

Albanèse, V., Yam, A. Y., Baughman, J., Parnot, C., & Frydman, J. (2006). Systems analyses reveal two chaperone networks with distinct functions in eukaryotic cells. *Cell 124*, 75-88.

Ali, M.M., Roe, S.M., Vaughan, C.K., Meyer, P., Panaretou, B., Piper, P.W., Prodromou, C., Pearl, L.H. (2006). Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. *Nature*. *440(7087),* 1013-7.

Altschul, S. F., Gish, W., Miller, W., Myers, E., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402.

Americo, T. A., Chiarini, L. B., & Linden, R. (2007). Signaling induced by hop/STI-1depends on endocytosis. *Biochemical and Biophysical Research Communications. 358*, 620-625.

Arkin, M. R., & Wells, J. a. (2004). Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature reviews. Drug discovery*, *3*(4), 301–17.

Bailey, T L, & Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics (Oxford, England)*, *14*(1), 48–54.

Bailey, Timothy L, Boden, M., Buske, F. a, Frith, M., Grant, C. E., Clementi, L., Ren, J., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server issue), W202–8.

Bailey, Timothy L., & Charles, E. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (pp. 28–36). Menlo Park, California: AAAI Press.

Baugh, E. H., Lyskov, S., Weitzner, B. D., & Gray, J. J. (2011). Real-time PyMOL visualization for Rosetta and PyRosetta. *PloS One*, *6*(8), e21931.

Beaumont, M. a, & Rannala, B. (2004). The Bayesian revolution in genetics. *Nature reviews. Genetics*, *5*(4), 251–61.

Benner, S. a. (2001). Natural progression. *Nature*, *409*(6819), 459.

Berman, H. M., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, *10*(12), 98.

Berman, J., Westbrook, Z., Feng, G., Gilliland, T. N., Bhat, H., Weissig, I. N., Shindyalov, P. E., et al. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*, 235–242.

Birkholtz, L., Joubert, F., Neitz, A. W. H., & Louw, A. I. (2003). Comparative properties of a three-dimensional model of Plasmodium falciparum ornithine decarboxylase. *Proteins*, *50*(3), 463–473.

Boniecki, M., Rotkiewicz, P., Skolnick, J., & Kolinski, A. (2003). Protein fragment reconstruction using various modeling techniques. *Journal of Computer-aided Molecular Design*, *17*(11), 725–38.

Brayton, K. a, Lau, A. O. T., Herndon, D. R., Hannick, L., Kappmeyer, L. S., Berens, S. J., Bidwell, S. L., et al. (2007). Genome Sequence of Babesia bovis and Comparative Analysis of Apicomplexan Hemoprotozoa. *PLoS Pathogens*, *3*(10), 1401–1013.

Charpian, S., Przyborski, J. M., & Strasse, K. V. F. (2008). Protein Transport Across the Parasitophorous Vacuole of Plasmodium falciparum: Into the Great Wide Open, *Traffic (8)*, 157–165.

Chaudhury, S., Lyskov, S., & Gray, J. J. (2010). PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics (Oxford, England)*, *26*(5), 689–91.

Chen, S., Prapapanich, V., Rimerman, R. A., Honoré, B., and Smith, D. F. (1996). Interactions of p60, a mediator of progesterone receptor assembly, with heat shock proteins Hsp90 and Hsp70. The Journal of *Molecular Endocrinology. 10,* 682-693.

Chen, S., and Smith, D. F. (1998). Hop as an adaptor in the heat shock protein 70 (Hsp70) and hsp90 chaperone machinery. *The Journal of Biological Chemistry. 273*, 35194-35200.

Cheung-Flynn, J., Roberts, P. J., Riggs, D. L., & Smith, D. F. (2003). C-terminal sequences outside the tetratricopeptide repeat domain of FKBP51 and FKBP52 cause differential binding to Hsp90. *The Journal of Biological Chemistry*, *278*(19), 17388–94.

Clare, J., Tate, S., Nobbs, M., & Romanos, M. (2000). Voltage-gated sodium channels as therapeutic targets. *Drug discovery today*, *5*(11), 506–520.

Coitinho, A. S., Lopes, M. H., Hajj, G. N. M., Rossato, J. I., Freitas, A. R., Castro, C. C., Cammarota, M., et al. (2007). Short-term memory formation and long-term memory consolidation are enhanced by cellular prion association to stress-inducible protein 1. *Neurobiology of Disease*, *26*(1), 282–290.

Corena, P., VanEkeris, L., Salazar, M. I., Bowers, D., Fiedler, M. M., Silverman, D., Tu, C., et al. (2005). Carbonic anhydrase in the adult mosquito midgut. *The Journal of Experimental Biology*, *208*(Pt 17), 3263–73.

Daniel, S., Bradley, G., Longshaw, V. M., Söti, C., Csermely, P., & Blatch, G. L. (2008). Nuclear translocation of the phosphoprotein Hop (Hsp70/Hsp90 organizing protein) occurs under heat shock, and its proposed nuclear localization signal is involved in Hsp90 binding. *Biochimica et Biophysica Acta*, *1783*(6), 1003–14.

DeLano, W.L. (2002)The PyMOL Molecular Graphics System. DeLano Scientific, San Carlos, CA, USA.

Deponte, M., Hoppe, H.C., Lee, M.C., Maier, A.G., Richard, D., Rug, M., Spielmann, T., Przyborski, J.M. (2012). Wherever I may roam: protein and membrane trafficking in *P. falciparum*-infected red blood cells. *Molecularand Biochemical Parasitology*, *186(2),* 95-116.

Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, *26*(8), 897–9.

Donnelly, A. C., Zhao, H., Reddy Kusuma, B., & Blagg, B. S. J. (2010). Cytotoxic sugar analogues of an optimized novobiocin scaffold. *Medical Chemistry Communications*, *1*(2), 165.

Eiben, C. B., Siegel, J. B., Bale, J. B., Cooper, S., Khatib, F., Shen, B. W., Stoddard, B. L., et al. (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature Biotechnology*, *130*, 190–192.

Eramian, D., Eswar, N., Shen, M.-Y., & Sali, A. (2008). How well can the accuracy of comparative protein structure models be predicted? *Protein Science*, *17*(11), 1881–93.

Erlich, R. B., Kahn, S. A., Avia, F. L., Martins, R. A. P., Linden, R., Chiarini, L. B., Martins, V. R., et al. (2007). STI1 Promotes Glioma Proliferation Through MAPK and PI3K Pathways. *GLIA*, *55*, 1690–1698.

Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., & Sali, A. (2008). Protein structure modeling with MODELLER. *Methods in Molecular Biology (Clifton, N.J.)*, *426*, 145–59.

Eustace B. K. & Jay D. G. (2004). Extracellular roles for the molecular chaperone, hsp90. *Cell Cycle, 3*, 1098-1100..

Figueireido, M. (2002). Unsupervised Learning of Finite Mixture Models. *Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 381–396.

Flom, G., Weekes, J., Williams, J. J., & Johnson, J. L. (2006). Effect of mutation of the tetratricopeptide repeat and asparatate-proline 2 domains of Sti1 on Hsp90 signaling and interaction in Saccharomyces cerevisiae. *Genetics 172*, 41-51.

Giordano, A., Whyte, P., Harlow, E., Jr., B. R. F., Beach, D., & Draetta, G. (1989). A 60 kd cdc2-associated polypeptide complexes with the E1A proteins in adenovirus-infected cells. *Cell*, *58*(5), 981 – 990.

Gitau, G. W., Mandal, P., Blatch, G. L., Przyborski, J., & Shonhai, A. (2012). Characterisation of the Plasmodium falciparum Hsp70-Hsp90 organising protein (PfHop). *Cell Stress & Chaperones*, *17*(2), 191–202.

Grigorieff, N. & Harrison, S.C. (2011). Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy. *Current Opinion in Structural Biology, 21(2)*, 265-73.

Gromov, P. S., & Celis, J. (1991). Identification of Two Molecular Chaperons (HSX70 , HSC70) in Mature Human Erythrocytes. *Experimental Cell Research*, *195*, 556–559.

Harrison, C. J., & Langdale, J. a. (2006). A step by step guide to phylogeny reconstruction. *The Plant Journal : for Cell and Molecular biology*, *45*(4), 561–72.

Hatherley, R. (2012). *In Silico Characterisation of the Four Canonical Plasmodium falciparum 70 kDa Heat Shock Proteins.* MScby Coursework / Thesis. Rhodes University.

Honoré, B., Leffers, H., Madsen, P., Rasmussen, H. H., Vandekerckhove, J., & Celis, J. E. (1992). Molecular cloning and expression of a transformingsensitive human protein containing the TPR motif and sharing identity to the stress-inducible yeast protein STI1. *The Journal of Biological Chemistry, 267,* 8485-8491.

Horibe, T., Kohno, M., Haramoto, M., Ohara, K., & Kawakami, K. (2011). Designed hybrid TPR peptide targeting Hsp90 as a novel anticancer agent. *Journal of Translational Medicine*, *9*(1), 8.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(Web Server issue), W5–9.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences*,*8*(3), 275–82.

Joosten, R. P., Joosten, K., Cohen, S. X., Vriend, G., & Perrakis, A. (2011). Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics (Oxford, England)*, *27*(24), 3392–8.

Joosten, R. P., Joosten, K., Murshudov, G. N., & Perrakis, A. (2012). PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallographica. Section D, Biological Crystallography*, *68*(4), 484–96.

Kajander, T., Sachs, J. N., Goldman, A., & Regan, L. (2009). Electrostatic interactions of Hsp-organizing protein tetratricopeptide domains with Hsp70 and Hsp90: computational analysis and protein engineering. *The Journal of Biological Chemistry*, *284*(37), 25364–74.

Kamionka, M., & Feigon, J. (2004). Structure of the XPC binding domain of hHR23A reveals hydrophobic patches for protein interaction. *Protein Science*, *13*, 2370–2377.

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.

Katoh, K. &Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics, 9,*276-285.

Kerfeld, C., & Scott, K. M.(2011). Using BLAST to teach "E-value-tionary" concepts. *PLoS Biology*, *9*(2), e1001014.

Kersey, P. J., Lawson, D., Birney, E., Derwent, P. S., Haimel, M., Herrero, J., Keenan, S., et al. (2010). Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Research*, *38*(Database issue), D563–9.

Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popovic, Z., Baker, D., et al. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(47), 18949–53.

Kortemme, T., & Baker, D. (2002). A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(22), 14116–21.

Kortemme, T., Kim, D. E., & Baker, D. (2004). Computational Alanine Scanning of Protein-Protein Interfaces. *Science STKE,* pl2.Krause, P. J. (2003). Babesiosis Diagnosis and Treatment. *Vector-borne and Zoonotic Diseases*, *3*(1), 45–51.

Krishnamoorthy, B., & Tropsha, a. (2003). Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, *19*(12), 1540–1548.

Külzer, S., Charnaud, S., Dagan, T., Riedel, J., Mandal, P., Pesce, E.R. Blatch, G.L., Crabb, B.S., Gilson, P.R. & Przyborski, J.M. (2012). *Plasmodium falciparum*-encoded exported hsp70/hsp40 chaperone/co-chaperone complexes within the host erythrocyte. *Cellular Microbiology, 14(11)*, 1784–1795.

Kumar, R., Musiyenko, A., & Barik, S. (2003). The heat shock protein 90 of *Plasmodium falciparum* and antimalarial activity of its inhibitor, geldanamycin. *Malaria Journal*, *2*, 30.

Kumar, S., Tamura, K., & Nei, M. (1994). MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Computational and Applied Bioscience*, *10*(2), 189 – 191.

Kunicki, J. B., Petersen, M. N., Alexander, L. D., Ardi, V. C., McConnell, J. R., & McAlpine, S. R. (2011). Synthesis and evaluation of biotinylated sansalvamide A analogs and their modulation of Hsp90. *Bioorganic & Medicinal Chemistry letters*, *21*(16), 4716–9.

Kyte, J., & Doolittle, R. F. (1982). A Simple Method for Displaying the Hydropathic Character of a Protein. *Journal of Molecular Biology*, *157*(1), 105 – 132.

Papadopoulos, J.S. & Agarwala, R.(2007).COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics, 23(9)*, 1073-1079.

Pare, J.M., Tahbaz, N., López-Orozco, J., LaPointe, P., Lasko, P. & Hobman, T.C. (2009). Hsp90 regulates the function of argonaute 2 and its recruitment to stress granules and P-bodies. *Molecular Biology of the Cell, 20(14)*, 3273-84.

Lassle, M., Blatch, G. L., Kundra, V., Takatori, T., & Zetter, B. R. (1997). Stress-inducible, Murine Protein mSTI1. *The Journal of Biological Chemistry*, *272*(3), 1876–1884.

Lazaridis, T., & Karplus, M. (1998). Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *Journal of Molecular Biology*, *288*(3), 477–87.

Lazaridis, T., & Karplus, M. (1999). Effective energy function for proteins in solution. *Proteins*, *35*(2), 133–52.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K. W., et al. (2011). Chapter nineteen – Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods in Enzymology*, *487*, 545–574.

Lee, C.-T., Graf, C., Mayer, F. J., Richter, S. M., & Mayer, M. P. (2012). Dynamics of the regulation of Hsp90 by the co-chaperone Sti1. *The EMBO Journal*, *90*, 1–11.

Li H. (2006) Constructing the TreeFam database. PhD thesis, the Institute of Theoretical Physics, Chinese Academy of Science.Li, H., Coghlan, A., Ruan, J., Coin, L. J., Hériché, J.-K., Osmotherly, L., Li, R., et al. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research*, *34*(Database issue), D572–80.

Li, H., & Durbin, R. (2007). Incorporating speices phylogeny in the reconstruction of gene trees. *Current Challenges and Problems in Phylogenetics*. An Isaac Newton Workshop Proceedings by The Wellcome Trust Sanger Institute.

Lima, F. R. S., Arantes, C. P., Muras, A. G., Nomizo, R., Brentani, R. R., & Martins, V. R. (2007). Cellular prion protein expression in astrocytes modulates neuronal survival and differentiation. *Journal of Neurochemistry*, *103*(6), 2164–76.

Lin, K., May, A. C. W. & Taylor, W. R. (2002). Threading using neural nEtwork (TUNE): the measure of protein sequence-structure compatibility., *Bioinformatics 18*(10), 1350–1357.

Longshaw, V. M., Chapple, J. P., Balda, M. S., Cheetham, M. E., & Blatch, G. L. (2004). Nuclear translocation of the Hsp70/Hsp90 organizing protein mSTI1 is regulated by cell cycle kinases. *Journal of Cell Science*, *117*(Pt 5), 701–10.

Lopes, M. H., Hajj, G. N., Muras, A. G., Mancini, G. L., Castro, R. M., Ribeiro, K. C., Brentani, R. R., Linden, R., and Martins, V. R. (2005). Interaction of cellular prion and stress-inducible protein 1 promotes neuritogenesis and neuroprotection by distinct signaling pathways. *Journal of Neuroscience, 25*, 11330-11339.

Lovell, S.C., Davis, I.W., Arendall, W.B. 3rd, de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C. (2003). Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*, *450*, 437–450.

Luthy, R., Bowie, J. U., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, *365*, 83–85.

Lyskov, S., & Gray, J. J. (2008). The RosettaDock server for local protein-protein docking. *Nucleic Acids Research*, *36*(Web Server issue), W233–8.

Ma, B., & Nussinov, R. (2007). Trp / Met / Phe Hot Spots in Protein-Protein Interactions: Potential Targets in Drug Design. *Current Topics in Medicinal Chemistry*, *7*(10), 999–1005.

Martins, V. R., Graner, E., Garcia-Abreu, J., de Souza, S. J., Mercadante, A. F., & Veiga, S. S., Zanata, S. M., Neto, V. M., and Brentani, R. R. (1997). Complementary hydropathy identifies a cellular prion protein receptor. *Nature Medicine*, *3*, 1376–1382.

MATLAB R2009a, The MathWorks Inc., Natrick, MA, 2009.

Matts, R. L., Dixit, A., Peterson, L. B., Sun, L., Voruganti, S., Kalyanaraman, P., Hartson, S. D., et al. (2011). Elucidation of the Hsp90 C-terminal inhibitor binding site. *ACS Chemical Biology*, *6*(8), 800–7.

Mehrpour, M., Esclatine, A., Beau, I. & Codogno, P. (2010). Overview of macroautophagy regulation in mammalian cells. *Cell Research, 20(7)*, 748-62.

Melo, F., & Feytmans, E. (1998). Assessing Protein Structures with a Non-local Atomic Interaction Energy. *Journal of Molecular Biology*, *277*, 1141–1152.

Misura, K. M. S., & Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. *Proteins*, *59*(1), 15–29.

Nicolet, C. M., & Craig, E. A. (1989). Isolation and Characterization of STIJ , a Stress-Inducible Gene from Saccharomyces cerevisiae, *9*(9), 3638–3646.

Odunuga, O. O., Longshaw, V. M., & Blatch, G. L. (2004). Hop: more than an Hsp70/Hsp90 adaptor protein. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *26*(10), 1058–68.

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). How many drug targets are there? *Nature Reviews. Drug Discovery*, *5*(12), 993–6.

Pasini, E. M., Kirkegaard, M., Salerno, D., Mortensen, P., Mann, M., & Thomas, A. W. (2008). Deep coverage mouse red blood cell proteome: a first comparison with the human red blood cell. *Molecular & Cellular Proteomics: MCP*, *7*(7), 1317–30.

Pawlowski, M., Gajda, M. J., Matlak, R., & Bujnicki, J. M. (2008). MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformatics*, *9*, 403.

Pei, J. (2008). Multiple protein sequence alignment. *Current opinion in structural biology*, *18*(3), 382–6.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2008). FastBLAST: homology relationships for millions of proteins. *PloS one*, *3*(10), e3589.

Pruitt, K. D., & Maglott, D. R. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, *29*(1), 137–40.

Roffé, M., Beraldo, F. H., Bester, R., Nunziante, M., Bach, C., Mancini, G., Gilch, S., et al. (2010). Prion protein interaction with stress-inducible protein 1 enhances neuronal protein synthesis via mTOR. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(29), 13147–52.

Romano, S. a, Cordeiro, Y., Lima, L. M. T. R., Lopes, M. H., Silva, J. L., Foguel, D., & Linden, R. (2009). Reciprocal remodeling upon binding of the prion protein to its signaling partner hop/STI1. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, *23*(12), 4308–16.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., Hériché, J.-K., et al. (2008). TreeFam: 2008 Update. *Nucleic Acids Research*, *36*(Database Issue), D735–D740.

Sakudo, A., Lee, D. C., Li, S., Nakamura, T., Matsumoto, Y., Saeki, K., Itohara, S., Ikuta, K., & Onodera, T. (2005). PrP^{c} cooperates with STI1 to regulate SOD activity in PrP^{c}-deficient neuronal cell line. *Biochemistry and Biophysics Research Communications, 328,* 14-19.

Sali, A., & Blundell, T. L. (1993). Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, *234*, 779–815.

Scheufler, C., Brinker, A., Bourenkov, G., Pegoraro, S., Moroder, L., Bartunik, H., Hartl, F. U., et al. (2000). Structure of TPR domain-peptide complexes: critical elements in the assembly of the Hsp70-Hsp90 multichaperone machine. *Cell*, *101*(2), 199–210.

Schmid, A. B., Lagleder, S., Gräwert, M. A., Röhl, A., Hagn, F., Wandinger, S. K., Cox, M. B., et al. (2012). The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop. *The EMBO Journal*, *31*, 1506–1517.

Schwieters, C.D., Kuszewski, J.J., Tjandra, N., Clore, G.M. (2003) The Xplor-Nih NMR molecular structure determination package. *Journal of Magnetic Resonance,160*, 65–73.

Searls, D. B. (2003). Pharmacophylogenomics: genes, evolution and drug targets. *Nature reviews. Drug Discovery*, *2*(8), 613–23.

Shen, M.-Y., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, *15*(11), 2507–24.

Shonhai, A. (2010). Plasmodial heat shock proteins: targets for chemotherapy. *FEMS Immunology and Medical Microbiology*, *58*(1), 61–74.

Sims, J. D., McCready, J., & Jay, D. G. (2011). Extracellular heat shock protein (Hsp)70 and Hsp90α assist in matrix metalloproteinase-2 activation and breast cancer cell migration and invasion. *PloS One*, *6*(4), e18848.

Siwach, P., Sengupta, S., Parihar, R., & Ganesh, S. (2011). Proline repeats, in cis- and trans-positions, confer protection against the toxicity of misfolded proteins in a mammalian cellular model. *Neuroscience Research*, *70*(4), 435–41.

Smith, C. (2003). Hitting the target. *Nature Reviews*, *422*, 341–347.

Smith, D. F., Sullivan, W. P., Marion, T. N., Zaitsu, K., Madden, B., Mccormick, D. J., & Toft, D. O. (1993). Identification of a 60-kilodalton stress-related protein, p60, which interacts with hsp90 and hsp70. *Molecular and Cellular Biology*, *13*(2), 869–76.

Southworth, D. R., & Agard, D. a. (2011a). Client-loading conformation of the Hsp90 molecular chaperone revealed in the cryo-EM structure of the human Hsp90:Hop complex. *Molecular Cell*, *42*(6), 771–81.

Southworth, D. R., & Agard, D. a. (2011b). Client-loading conformation of the Hsp90 molecular chaperone revealed in the cryo-EM structure of the human Hsp90:Hop complex. *Molecular Cell*, *42*(6), 771–81.

Subramanian, A. R., Kaufmann, M., & Morgenstern, B. (2008). DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms for Molecular Biolog* , *3*, 6.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5 : Molecular Evolutionary Genetics Analysis Using Maximum Likelihood , Evolutionary Distance , and Maximum Parsimony Methods. *Molecular Biology and Evolution*, *28*(10), 2731–2739.

Tastan Bishop, A. O., De Beer, T. A. P., & Joubert, F. (2008). Protein homology modelling and its use in South Africa. *South African Journal Of Science*, *104*, 2–6.

Tastan Bishop, Ö., & Kroon, M. (2011). Study of protein complexes via homology modeling , applied to cysteine proteases and their protein inhibitors. *Journal of Molecular Modelling*. 17, 3163 - 3172.

Tilley, L., Dixon, M. W. , & Kirk, K. (2011). The *Plasmodium falciparum*-infected red blood cell. *The International Journal of Biochemistry & Cell Biology*, *43*(6), 839–42.

Tina, K. G., Bhadra, R., & Srinivasan, N. (2007). PIC: Protein Interactions Calculator. *Nucleic Acids Research*, *35*(Web Server issue), W473–6.

Tyka, M. D., Keedy, D. A., André, I., Dimaio, F., Song, Y., Richardson, D. C., Richardsonb, J. S., et al. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology*, *405*(2), 607–618.

van Der Spuy, J., Kana, B. D., Dirr, H. W., & Blatch, G. L. (2000). Heat shock cognate protein 70 chaperone-binding site in the co- chaperone murine stress inducible protein 1 maps to within three consecutive tetratricopeptide repeat motifs. *Biochemical Journal, 345,* 645-651.

Vyas, V.K., Ukawala, R.D., Ghate, M. &Chintha, C. (2012) Homology modeling a fast tool for drug discovery: Current perspectives. *The Indian Journal of Pharmaceutical Sciences, 74(1)*, 1-17.

Wallner, B., & Elofsson, A. (2006). Identification of correct regions in protein models using structural , alignment, and consensus information. *Protein Science*, *15*, 900–913.

Wang, T.-H., Chao, A., Tsai, C.-L., Chang, C.-L., Chen, S.-H., Lee, Y.-S., Chen, J.-K., et al. (2010). Stress-induced phosphoprotein 1 as a secreted biomarker for human ovarian cancer promotes cancer cell proliferation. *Molecular &Cellular Proteomics*, *9*(9), 1873–84.

Weiss, M. J., & Dos Santos, C. O. (2009). Chaperoning erythropoiesis. *Blood*, *113*(10), 2136–44.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, *18*(5), 691–9.

WHO: Global Malaria Programme. (2011). *World malaria report: 2011* (p. 248). Geneva.

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.

Williamson, M. P. (1994). The structure and function of proline-rich regions in proteins. *The Biochemical Journal*, *297(Pt 2)*, 249–60.

Wu, S., & Zhang, Y. (2007). LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, *35*(10), 3375–82.

Yi, F., Zhu, P., Southall, N., Inglese, J., Austin, C. P., Zheng, W., & Regan, L. (2009). An AlphaScreen-based high-throughput screen to identify inhibitors of Hsp90-cochaperone interaction. *Journal of Biomolecular Screening*, *14*(3), 273–81.

Zanata, S. M., Lopes, M. H., Mercadante, A. F., Hajj, G. N. M., Chiarini, L. B., Nomizo, R., Freitas, A. R. O., et al. (2002). Stress-inducible protein 1 is a cell surface ligand for cellular prion that triggers neuroprotection. *EMBO*, *21*(13), 3307–3316.

Zemla, A., Venclovas, Č., Moult, J., & Fidelis, K. (2002). Processing and evaluation of predictions in CASP4. *Proteins*, *45*(5), 13–21.

Zhang, T., Li, Y., Yu, Y., Zou, P., Jiang, Y., & Sun, D. (2009). Characterization of celastrol to inhibit hsp90 and cdc37 interaction. *The Journal of Biological Chemistry*, *284*(51), 35381–9.

Appendix 1: Blast Results

Table A1.1: searching for Hop homologs on NCBI pBLAST's Genome Viewer with the protein sequence for gene entry PF3D7_1434300 from PlasmoDB.org

Species	Protein ID	E-value	Score	% Identity	Positives	Gaps	Gene ID	Description
Callithrix jacchus Length= 586	XP_002755509.1	2e-107	336 bits (861)	212/565 (38%)	326/565 (58%)	28/565 (5%)	GENE ID: 100409178 STIP1	Predicted stress-induced-phosphoprotein 1 isoform 1
Homo sapiens Length= 565	NP_006810.1	2e-106	333 bits (855)	212/565 (38%)	326/565 (58%)	28/565 (5%)	GENE ID: 10963 STIP1	stress-induced-phosphoprotein 1
Pan troglodytes Length= 560	XP_001163388.1	3e-105	332 bits (851)	211/560 (38%)	324/560 (58%)	28/560 (5%)	GENE ID: 451286 STIP1	PREDICTED: stress-induced-phosphoprotein 1 isoform 1
Rattusnorvegicus Length= 565	NP_620266.1	1e-107	337 bits (864)	212/565 (38%)	327/565 (58%	28/565 (5%)	GENE ID: 192277 Stip1	stress-induced-phosphoprotein 1
Musmusculus Length= 565	NP_058017.1	3e-108	338 bits (868	214/565 (38%)	326/565 (58%)	28/565 (5%)	GENE ID: 20867 Stip1	stress-induced-phosphoprotein 1
Cricetulusgriseus Length= 565	NP_001233607.1	2e-108	338 bits (868)	214/565 (38%	325/565 (58%)	28/565 (5%)	GENE ID: 100689413 Stip1	stress-induced-phosphoprotein 1
Ornithorhynchus anatinus Length= 547	XP_001511150.1	1e-101	320 bits (819)	204/547 (37%)	308/547 (56%)	40/547 (7%)	GENE ID: 100080264 STIP1	PREDICTED: stress-induced-phosphoprotein 1, partial
Ailuropodamelan oleuca Length= 565	XP_002916718.1	8e-108	337 bits (864)	214/565 (38%)	323/565 (57%)	28/565 (5%)	GENE ID: 100475133 LOC100475133	PREDICTED: stress-induced-phosphoprotein 1-like
Bostaurus Length= 565	NP_001030569.1	5e-107	335 bits (859)	214/565 (38%)	324/565 (57%)	28/565 (5%)	GENE ID: 617109 STIP1	stress-induced-phosphoprotein 1
Sus scrofa Length= 565	XP_003353842.1	3e-108	338 bits (866)	216/565 (38%)	323/565 (57%)	28/565 (5%)	GENE ID: 100623923 LOC100623923	PREDICTED: stress-induced-phosphoprotein 1-like
Xenopus (Silurana) tropicalis Length= 573	NP_989360.1	3e-108	338 bits (867)	217/573 (38%)	325/573 (57%)	44/573 (8%)	GENE ID: 394990 stip1	stress-induced-phosphoprotein 1
Oreochromisnilot icus Length= 571	XP_003450486.1	8e-111	345 bits (884)	215/571 (38%)	328/571 (57%)	41/571 (7%)	GENE ID: 100696373 LOC100696373	stress-induced-phosphoprotein 1-like
Daniorerio Length= 565	NP_001007767.1	3e-109	341 bits (874)	211/565 (37%)	325/565 (58%)	29/565 (5%)	GENE ID: 493606 stip1	stress-induced-phosphoprotein 1

Anoliscarolinensis Length= 566	XP_003228007.1	1e-112	349 bits (896)	216/566 (38%)	327/566 (58%)	30/566 (5%)	GENE ID: 100563364 LOC100563364	PREDICTED: stress-induced-phosphoprotein 1-like
Taeniopygia guttata Length= 528	XP_002198951.1	5e-18	86.7	35%	Coverage = 84%		GENE ID: 100229922 LOC100229922	PREDICTED: RNA polymerase II associated protein 3, partial
Meleagrisgallopavo Length 665	XP_003202082.1	5e-19	90.1	33%	Coverage = 76%		GENE ID: 100541433 LOC100541433	PREDICTED: LOW QUALITY PROTEIN: RNA polymerase II-associated protein 3-like
Gallus gallus Length 665	XP_418360.3	7e-17	84.0	35%	Coverage = 84%		GENE ID: 420249 SPAG1	PREDICTED: sperm-associated antigen 1
Monodelphisdomestica Length=441 (Marsupial)	XP_001372931.2	3e-15	78.2	29%	Coverage = 90%		GENE ID: 100020424 LOC100020424	PREDICTED: small glutamine-rich tetratricopeptide repeat-containing protein alpha-like
Acyrthosiphon pisum Length=565	XP_001950745.1	4e-112	348 bits (892)	204/565 (36%)	319/565 (56%)	31/565 (5%)	GENE ID: 100167947 LOC100167947	PREDICTED: stress-induced-phosphoprotein 1-like
Anopheles gambiae due to problems with alignment not used	XP_319365.4	4e-80	257	41%	Coverage = 80%		GENE ID: 1279608 AgaP_AGAP010188	AGAP010188-PA [Anopheles gambiae str. PEST]
Bombusterrestris Length=565	XP_003402501.1	8e-114	352 bits (902)	207/555 (37%)	324/555 (58%)	24/555 (4%)	GENE ID: 100631059 cactus-2	PREDICTED: hypothetical protein LOC100631059
Drosophila melanogaster Length = 490	NP_477354.1	248	1e-74	51%	86%		GENE ID: 33202 Hop	Hsp70/Hsp90 organizing protein homolog
Nasoniavitripennis Length=565	XP_001603429.1	1e-106	333 bits (855)	204/569 (36%)	323/569 (57%)	32/569 (6%)	GENE ID: 100119701 LOC100119701	PREDICTED: stress-induced-phosphoprotein 1-like
Amphimedonqueenslandica Length=554	XP_003387638.1	8e-101	318 bits (815)	198/564 (35%)	319/564 (57%)	20/564 (4%)	GENE ID: 100633434 LOC100633434	PREDICTED: Stress-induced phosphoprotein 1-like

Cionaintestinalis Length=570	XP_002128875.1	2e-100	317 bits (813)	197/570 (35%)	314/570 (55%)	37/570 (6%)	GENE ID: 100181490 LOC100181490	PREDICTED: similar to Stress-induced-phosphoprotein 1 (STI1) (Hsc70/Hsp90-organizing protein) (Hop)
Saccoglossuskowalevskii Length=310	XP_002733893.1	3e-75	244 bits (622)	127/315 (40%)	203/315 (64%)	8/315 (3%)	GENE ID: 100374768 LOC100374768	PREDICTED: stress-induced-phosphoprotein 1 (Hsp70/Hsp90-organizing protein)-like
Hydra magnipapillata Length=534	XP_002160503.1	8e-97	307 bits (787)	195/560 (35%)	308/560 (55%)	28/560 (5%)	GENE ID: 100203295 LOC100203295	PREDICTED: similar to stress-induced-phosphoprotein 1 (Hsp70/Hsp90-organizing protein), partial
Caenorhabditiselegans Length=320	NP_503322.1	2e-73	240 bits (612)	126/316 (40%)	205/316 (65%)	8/316 (3%)	GENE ID: 178587 sti-1	Protein STI-1
Caenorhabditisbriggsae Length=320	XP_002634443.1	5e-72	236 bits (601)	126/316 (40%)	203/316 (64%)	Gaps = 8/316 (3%)	GENE ID: 8576439 CBG04457	Hypothetical protein CBG04457
Babesiabovis Length=546	XP_001611358.1	0.0	546 bits (1407)	273/553 (49%)	394/553 (71%)	12/553 (2%)	GENE ID: 5479603 BBOV_III002230	tetratricopeptide repeat domain containing protein
Cryptosporidium parvum [Iowa II] Length=326	XP_001388209.1	6e-131	387 bits (994)	183/318 (58%)	251/318 (79%)	0/318 (0%)	GENE ID: 3373446 cgd2_1850	stress-induced protein sti1-like protein
Leishmaniabraziliensis[MHOM/BR/75/M2904] Length=547	XP_001562145.1	9e-110	341 bits (875)	209/563 (37%)	320/563 (57%)	26/563 (5%)	GENE ID: 5413050 LBRM_08_0880	stress-induced protein sti1
Leishmania major[strain Friedlin] Length=546	XP_001681140.1	7e-118	362 bits (929)	214/563 (38%)	326/563 (58%)	27/563 (5%)	GENE ID: 5649395 LMJF_08_1110	stress-induced protein sti1
Leishmaniainfantum[JPCM5] Length=546	XP_001463435.1	2e-118	363 bits (933)	215/563 (38%)	326/563 (58%)	27/563 (5%)	GENE ID: 5066714 LINJ_08_1020	stress-induced protein sti1
Plasmodium berghei[strain ANKA] Length=559	XP_677465.1	0.0	880 bits (2275)	462/564 (82%)	519/564 (92%)	5/564 (1%)	GENE ID: 3426000 PB000909.03.0	hypothetical protein
Plasmodium chabaudichabaudi Length=559	XP_745506.1	0.0	880 bits (2273)	460/564 (82%)	519/564 (92%)	5/564 (1%)	GENE ID: 3498629 PC000814.02.0	hypothetical protein

Plasmodium falciparum [3D7] Length=564	XP_001348498.1	0.0	1144 bits (2959)	564/564 (100%)	564/564 (100%)	0/564 (0%)	GENE ID: 811906 PF14_0324	Hsp70/Hsp90 organizing protein, putative
Theileriaparva[strain Muguga] Length=540	XP_763615.1	0.0	532 bits (1371)	286/554 (52%)	393/554 (71%)	18/554 (3%)	GENE ID: 3499840 TP03_0587	hypothetical protein
Trypanosomabrucei brucei[strain 927/4 GUTat10.1] Length=550	XP_844966.1	4e-110	342 bits (877)	205/562 (36%)	316/562 (56%)	21/562 (4%)	GENE ID: 3657403 Tb927.5.2940	stress-induced protein sti1
Plasmodium yoelii yoelii[17XNL] Length=559	XP_731105.1	0.0	882 bits (2278	463/564 (82%)	520/564 (92%)	5/564 (1%)	GENE ID: 3830331 PY03138	stress-induced protein Sti1
Aspergillusniger[CBS 513.88] Length=580	XP_001395168.2	1e-83	273 bits (699)	187/584 (32%)	304/584 (52%)	34/584 (6%)	GENE ID: 4985429 ANI_1_116104	heat shock protein STI1
Saccharomyces cerevisiae[S288c] Length=589	NP_014670.1	2e-97	310 bits (793)	203/593 (34%)	316/593 (53%)	46/593 (8%)	GENE ID: 854192 STI1	Sti1p
Oryzasativa[Japonica Group] Length=578	NP_001047563.1	8e-116	358 bits (920)	213/600 (36%)	328/600 (55%)	65/600 (11%)	GENE ID: 4330134 Os02g0644100	Os02g0644100
Arabidopsis thaliana Length=558	NP_001031620.1	5e-121	372 bits (954)	218/572 (38%)	334/572 (58%)	29/572 (5%)	GENE ID: 826849 AT4G12400	putative stress-inducible protein
Aspergillusclavatus[NRRL 1] Length=581	XP_001272361.1	1e-81	268 bits (686)	188/585 (32%)	297/585 (51%)	35/585 (6%)	GENE ID: 4704565 ACLA_065690	heat shock protein (Sti1), putative
Trichophytonrubrum[CBS 118892] Length=578	XP_003232880.1	2e-88	286 bits (731)	202/585 (35%)	299/585 (51%)	37/585 (6%)	GENE ID: 10372247 TERG_06870	heat shock protein STI1
Theileriaannulata [strain Ankara] Length=540	XP_955292.1	0.0	543 bits (1398)	280/554 (51%)	393/554 (71%)	18/554 (3%)	GENE ID: 3865063 TA18515	hypothetical protein, conserved
Neosporacaninum[Liverpool] Length=563	XP_003880293.1	2e-168	493 bits (1269)	268/566 (47%)	370/566 (65%)	7/566 (1%)	GENE ID: 13446323 NCLIV_007330	similar to uniprot|P15705 Saccharomyces cerevisiae YOR027w STI1, related
Neosartoryafischeri[NRRL 181] Length=582	XP_001262823.1	1e-90	292 bits (747)	191/582 (33%)	299/582 (51%)	29/582 (5%)	GENE ID: 4589462 NFIA_114590	heat shock protein (Sti1), putative
Talaromycesstipitatus[ATCC 10500] Length=577	XP_002478770.1	7e-83	272 bits (695)	184/586 (31%)	296/586 (51%)	41/586 (7%)	GENE ID: 8108134 TSTA_090460	heat shock protein (Sti1), putative

Penicilliummarneffei[ATCC 18224] Length=578	XP_002146473.1	4e-84	275 bits (703)	182/590 (31%)	297/590 (50%)	48/590 (8%)	GENE ID: 7024148 PMAA_070140	heat shock protein (Sti1), putative
Paracoccidioides brasiliensis[possibly new species; '*lutzii*', Pb01] Length=578	XP_002791265.1	2e-92	298 bits (763)	199/579 (34%)	299/579 (52%)	Gaps = 26/579 (4%)	GENE ID: 9094424 PAAG_06811	heat shock protein STI1
Schizosaccharomycesjaponicus[yFS275] Length=582	XP_002174852.1	1e-96	305 bits (781)	200/585 (34%)	313/585 (54%)	33/585 (6%)	GENE ID: 7052233 SJAG_03717	chaperone activator Sti1
Candida tropicalis[MYA-3404] Length=579	XP_002551007.1	2e-97	309 bits (792)	198/585 (34%)	310/585 (53%	36/585 (6%)	GENE ID: 8299626 CTRG_05305	heat shock protein STI1
Saccharomyces cerevisiae[S288c] Length=589 *Referred to by Schmid et al., 2012 as strain YOR0W287	NP_014670.1	2e-96	310 bits (793)	203/593 (34%)	316/593 (53%)	46/593 (8%)	GENE ID: 854192 STI1	Sti1p
Pyrenophoratriticirepentis[Pt-1C-BFP] Length=576	XP_001940455.1	6e-87	282 bits (722)	189/581 (33%)	308/581 (53%)	Gaps = 32/581 (6%)	GENE ID: 6348424 PTRG_10123	heat shock protein STI1
Arthrodermagypseum[CBS 118893] Length=578	XP_003175251.1	5e-89	288 bits (736)	202/589 (34%)	303/589 (51%)	45/589 (8%)	GENE ID: 10030557 MGYG_02781	heat shock protein STI1
Candida albicans[SC5314] Length=590	XP_714740.1	4e-88	286 bits (732)	198/598 (33%)	304/598 (51%	51/598 (9%)	GENE ID: 3643631 STI1	hypothetical protein CaO19.10702
Plasmodium knowlesi[strain H] Length=560	XP_002260669.1	0.0	885 bits (2287)	460/564 (82%)	509/564 (90%)	4/564 (1%)	GENE ID: 7322649 PKH_131500	hypothetical protein, conserved in Apicomplexan species
Verticilliumalbo-atrum[VaMs.102] Length=584	XP_002999908.1	4e-87	283 bits (724)	195/586 (33%)	296/586 (51%)	36/586 (6%)	GENE ID: 9531727 VDBG_09948	heat shock protein STI1
Lodderomyceselongisporus [NRRL YB-4239] Length=596	XP_001524727.1	8e-88	285 bits (728)	196/600 (33%)	314/600 (52%)	49/600 (8%)	GENE ID: 5232040 LELG_03759	heat shock protein STI1
Schizosaccharomycespombe [972h-] Length=591	NP_588123.1	3e-92	296 bits (757)	202/596 (34%)	302/596 (51%)	46/596 (8%)	GENE ID: 2539474 sti1	chaperone activator Sti1 (predicted)

Aspergillusoryzae [RIB40] Length=579	XP_001825463.1	9e-87	285 bits (728)	195/584 (33%)	300/584 (51%)	35/584 (6%)	GENE ID: 5997558 AOR_1_950074	shock protein STI1
Aspergillusfumigatus[A1163] Length=585	XP_746746.1	3e-84	278 bits (711)	183/559 (33%)	287/559 (51%)	29/559 (5%)	GENE ID: 3504281 AFUA_7G01860	heat shock protein (Sti1)
Aspergillusflavus[NRRL3357] Length=579	XP_002380660.1	1e-87	285 bits (728)	195/584 (33%)	300/584 (51%)	35/584 (6%)	GENE ID: 7914463 AFLA_071010	heat shock protein (Sti1), putative

Table A1.2: searching for Hsp90 homologs on NCBI pBLAST's Genome Viewer with the protein sequence from SchmidCYS yeast's Hsp90 Structure

Species	Seq ID	E-value	Score	% Identity	Positives	Gaps	Gene ID	Description
Plasmodium falciparum [3D7] Length=927	XP_001348591.1	2e-78	261 bits (668)	147/409 (36%)	255/409 (62%)	18/409 (4%)	GENE ID: 811999 PF14_0417	HSP90
Homo sapiens Length=854	NP_001017963.2	4e-164	488 bits(1256)	244/399 (61%)	312/399 (78%)	2/399 (1%)	GENE ID: 3320 HSP90AA1	heat shock protein HSP 90-alpha isoform 1 (cytosolic)
Caenorhabditiselegans Length=702	NP_506626.1	2e-165	485 bits (1248)	246/399 (62%)	308/399 (77%)	3/399 (1%)	GENE ID: 179971 daf-21	Protein DAF-21 (Abnormal Dauer Formation protein 21)

Table A1.3: searching for Hsp90 homologs on NCBI pBLAST's Genome Viewer with the protein sequence from XP_001348591.1 (PfHsp90, from Table A1.2)

Species	Seq ID	E-value	Score	% Identity	Positives	Gaps	Gene ID	Description
Caenorhabditiselegans Length=702	NP_506626.1	2e-79	261 bits (668)	143/410 (35%)	251/410 (61%)	13/410 (3%)	GENE ID: 179971 daf-21	Protein DAF-21 (Abnormal Dauer Formation protein 21)
Homo sapiens Length=854	NP_001017963.2	3e-76	258 bits (658)	146/410 (36%)	246/410 (60%)	12/410 (3%)	GENE ID: 3320 HSP90AA1	heat shock protein HSP 90-alpha isoform 1 (cytosolic)

Appendix 2: Input Code and Scripts

Section 1: Checking Pairwise Sequence Identity Between Each Species in a Large Subset of Data.

The following program is a simple checker program written solely for the purpose of checking the minimum sequence identity score in a large subset of pairwise sequence data. The data is stored in file that was produced by Jalview (see Pairwise_identities.fasta in the supplementary data on disk) when calculating the pairwise identity of each of the 60 sequences that have been locally aligned to each other sequence in the database used in the project (See Appendix 4).

This checker file is necessary as the pairwise sequence identity data file can become too large to search manually when the sequence dataset becomes very large. For example, in this project 60 species sequences were downloaded for use in the multiple sequence analysis. To perform a pairwise alignment analysis between all of the sequences in Jalview would result in 60 x 60 = 3600 pairwise alignments. Because Hop has a sequence length greater than 250 residues (approximately 520 residues), it was desirable to ascertain whether all sequences used in the multiple sequence alignment and phylogenetic studies in Chapter 2 could have plausibly been homologs, simply by showing that all sequences used in the analysis shared at least 25% sequence identity to all other sequences, putting them in the "Daylight" zone of homology detection (Rost, 1999; Venclovas, 2012).

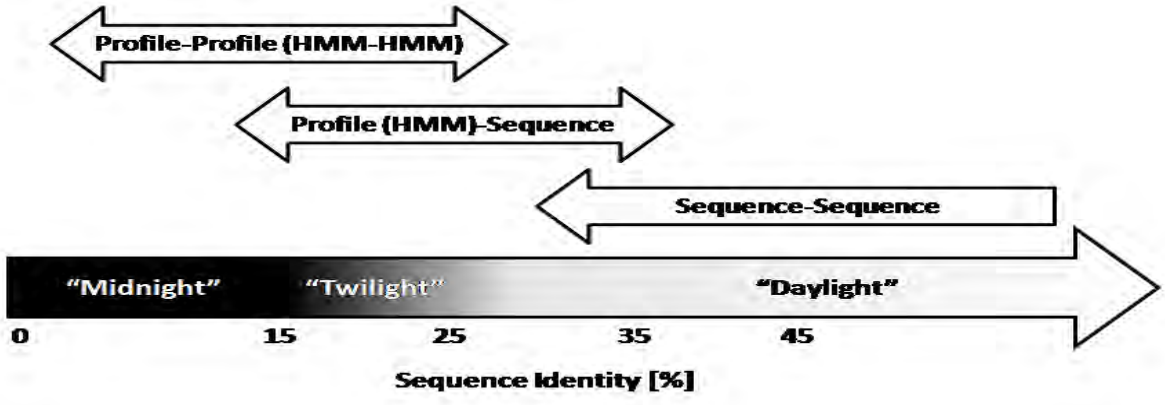


Figure A2.1: Adapted from Venclovas, 2012, Chapter 3, page 58. Detecting Homology is done through three methods of increasing complexity level, depending on the degree of sequence identity.

Finding minimum sequence identity in a large dataset: **pairwise_id_reporter.py**.

This script was run from the command line as follows:

[user@home]$ python pairwise_id_reporter.py

```
# This is a function that will parse the jalview pairwise identity
file (.fasta) and add axception where a value id below the user-
specified sequence_ID.

defpair_id_report(FAST,x):

      start = open(FAST, "r")
      newfile = start.read()
      stringlist = []
      stringlist = newfile.split("\n\nScore = ")

      idlist = []
      for k in range(len(stringlist)):
            stringlist[k] = stringlist[k].split("\n\nPercentage ID =
")

      for i in range(len(stringlist)):
            idlist.append(stringlist[i][1])

      idlist[len(idlist)-1].split("\n\n\n\n")
      for l in range(len(idlist)):
            idlist[l] = float(idlist[l])

      checker = 0
      for check in range(len(idlist)):
            if x >idlist[check]:
                  checker += 1
            else:
                  checker = checker

      if checker > 0:
            return "%i out of %i alignment identities are below
Percentage Identity Threshold %i%%" %(checker,len(idlist),x)
      else:
            return "All of %i alignment identities are above
Percentage Identity Threshold %i%%" %(len(idlist),x)

# Here the user inserts the file name to be searched and the minimum
Pairwise Identity level sought.

printpair_id_report('Pairwise_identities.fasta',25)

#Script written by Crystal-Leigh Clitheroe (03-03-2012).
```

Section 2: Creating Alignment Files (.pir Extensions) and Modeling Scripts (.py Extensions) for Modeller.

Modeller requires three important files to create models; an alignment file (.pir), a template file (.pdb) and a command script that runs Modeller functions using Python code (.py), all within the same folder. In Section A of this appendix the production of models from three types of templates is discussed and file and modeling sample scripts are provided. All modeling scripts are based on and modified from templates provided in the Modeller manual.

2-1) The first example script is for creating a model from a single chain template.

Example for a single chain alignment file: **DP1.pir**

```
>P1;2LLV_average_control
sequence:2LLV_average_control:1:A:70::: : 0.00: 0.00
QPDLGLTQLFADPNLIENLKKNPKT
SEMMKDPQLVAKLIGYKQN-PQAIGQDLFTDPRLMTIMATLMGVDLN*

>P1;2LLV_average
structureX:2LLV_average:127:A: 197::: : 0.00: 0.00
QPDLGLTQLFADPNLIENLKKNPKT
SEMMKDPQLVAKLIGYKQN-PQAIGQDLFTDPRLMTIMATLMGVDLN*
```

The template and target sequence must always align exactly and the alignment ends with "*" for both sequences. Only standard amino acids can be used in Modeller so if the template model contains non-standard amino acids, one will have to try change them to their standard forms (if possible, it is better to write to the authors of this template for their advice on modifications) or find another template.

Example of modeling script for single chain file using template 2LLV_average.pdb:
single_chain_make.py

This script was run from the command line as follows:
[user@home]$ python single_chain_make.py

```
# This script requires Modeller version 9.10 to run
# Homology modeling by the automodel and MyModel class

frommodeller import *                  # Load standard Modeller
classes
frommodeller.automodel import *  # Load the automodel class

log.verbose()          # request verbose output
env = environ()        # create new MODELLER environment to build the
models

# directories for input atom files
env.io.atom_files_directory = '.'

a = automodel(env,
alnfile = 'DP1.pir',                   # alignment filename
knowns = '2LLV_average',               # codes of the templates
sequence = '2LLV_average_control')     # code of the target

#Index Model/s
a.starting_model= 1
a.ending_model = 100


# Thorough MD optimization:
a.md_level = refine.very_slow

a.make()
```

Notice the names of the template structure and the target sequence is identical in each file. Notice also the blue highlighting in the DP1.pir file. For the template, the numbering of the residues can merely be 1 no matter how long the model chain will be. However the template numbering must reflect the exact numbering in the template.pdb file. This numbering will change if one trims the ends of their template.

2-2) The next example script is for creating a model from a multi chain template.

Example for a multi-chain alignment file: **3uq3_TPR2ab.pir**

```
>P1;3uq3_TPR2ab_control
sequence:3uq3_TPR2ab_control:1:A:249:B:C: : 0.00: 0.00
ADKEKAEGNKFYKARQFDEAIEHYNKAWELH-KDITYLNNRAAAEYEKGEYETAI
STLNDAVEQGREMRADYKVISKSFARIGNAYHKLGDLKKTIEYYQKSLTEHRTADILTKL
RNAEKELKKAEAEAYVNPEKAEEARLEGKEYFTKSDWPNAVKAYTEMIKRAPEDARGYSN
RAAALAKLMSFPEAIADCNKAIEKDPNFVRAYIRKATAQIAVKEYASALETLDAARTKDA
EVNNGSSAR/MEEVD/EVD*

>P1;3uq3_TPR2ab
structureX:3uq3_TPR2ab:262:A: 504:B:C: : 0.00: 0.00
ADKEKAEGNKFYKARQFDEAIEHYNKAWELH-KDITYLNNRAAAEYEKGEYETAI
STLNDAVEQGREMRADYKVISKSFARIGNAYHKLGDLKKTIEYYQKSLTEHRTADILTKL
RNAEKELKKAEAEAYVNPEKAEEARLEGKEYFTKSDWPNAVKAYTEMIKRAPEDARGYSN
RAAALAKLMSFPEAIADCNKAIEKDPNFVRAYIRKATAQIAVKEYASALETLDAARTKDA
EVNNGSSAR/MEEVD/EVD*
```

Notice again the regions in the .pir file that are highlighted. Each of the chains is separated by a backslash and the new chain names are inserted between the colon symbols following the last residue number of chain A. There is no need to number the residues for subsequent chains. Other than that, the modeling script that accompanies the multi-chain template and model takes an identical format to that of the single-chain script:

Example of modeling script for single chain file using template 3uq3_TPR2ab.pdb:
multi_chain_make.py

This script was run from the command line as follows:
[user@home]$ python multi_chain_make.py

```
# This script requires Modeller version 9.10 to run
# Homology modeling by the automodel and MyModel class

frommodeller import *                  # Load standard Modeller
classes
frommodeller.automodel import *  # Load the automodel class

log.verbose()               # request verbose output
env = environ()             # create a new MODELLER environment to
build in

# directories for input atom files
env.io.atom_files_directory = '.'


a = automodel(env,
alnfile = 'PfTPR2ab.pir',                    # alignment filename
knowns = '3uq3_TPR2ab',                      # codes of the templates
sequence = '3uq3_TPR2ab_control')            # code of the target

#Index Model/s
a.starting_model= 1
a.ending_model = 100

# Thorough MD optimization:
a.md_level = refine.very_slow

a.make()
```

2-3) The next example script is for creating a model from multiple, multi chain templates.

Example for a multi-template alignment file: **3uq3_TPR2ab_multitemp.pir**

```
>P1;PfTPR2ab2yeast
sequence:PfTPR2ab2yeast:1:A:249:B:C: : 0.00: 0.00
ADKEKAEGNKFYKARQFDEAIEHYNKAWELH-KDITYLNNRAAAEYEKGEYETAI
STLNDAVEQGREMRADYKVISKSFARIGNAYHKLGDLKKTIEYYQKSLTEHRTADILTKL
RNAEKELKKAEAEAYVNPEKAEEARLEGKEYFTKSDWPNAVKAYTEMIKRAPEDARGYSN
RAAALAKLMSFPEAIADCNKAIEKDPNFVRAYIRKATAQIAVKEYASALETLDAARTKDA
EVQQGSSAR/MEEVD/PTVEEVD*

>P1;3uq3_TPR2ab
structureX:3uq3_TPR2ab:262:A: 504:B:C: : 0.00: 0.00
ADKEKAEGNKFYKARQFDEAIEHYNKAWELH-KDITYLNNRAAAEYEKGEYETAI
STLNDAVEQGREMRADYKVISKSFARIGNAYHKLGDLKKTIEYYQKSLTEHRTADILTKL
RNAEKELKKAEAEAYVNPEKAEEARLEGKEYFTKSDWPNAVKAYTEMIKRAPEDARGYSN
RAAALAKLMSFPEAIADCNKAIEKDPNFVRAYIRKATAQIAVKEYASALETLDAARTKDA
EVNNGSSAR/MEEVD/----EVD*

>P1;3UPV_mod
structureX:3UPV_mod:261:A: 370:B:: : 0.00: 0.00
----------------------------------------------------------
------------------------------------------------------------
-------------------KAEEARLEGKEYFTKSDWPNAVKAYTEMIKRAPEDARGYSN
RAAALAKLMSFPEAIADCNKAIEKDPNFVRAYIRKATAQIAVKEYASALETLDAARTKDA
EVNNGSSAR/-----/PTVEEVD*
```

Notice again the regions in the .pir file that are highlighted. First, even though the second template does not possess the MEEVD chain, there is still a space for it in the alignment file. That is because all templates must be perfectly aligned to the target sequence, so any residues that either template does not possess, but exists in either the other templates or the target is represented by the gap symbol "-".

Example of modeling script for single chain file using templates 3uq3_TPR2ab.pdb and 3UPV_mod: **multi_chain_and_template_make.py**

This script was run from the command line as follows:

[user@home]$ python multi_chain_and_template_make.py

```
# This script requires Modeller version 9.10 to run
# Homology modeling by the automodel and MyModel class

frommodeller import *                  # Load standard Modeller
classes
frommodeller.automodel import *        # Load the automodel class

log.verbose()                     # request verbose output
env = environ()                   # create a new MODELLER environment


# directories for input atom files
env.io.atom_files_directory = '.'


a = automodel(env,
alnfile = 'PfTPR2ab_multitemp.pir',              # alignment
filename
knowns = ('3uq3_TPR2ab', '3UPV_mod'),            # codes of the
templates
sequence = 'PfTPR2ab2yeast')                     # code of the
target

#Index Model/s
a.starting_model= 2
a.ending_model = 100

# Thorough MD optimization:
a.md_level = refine.very_slow

a.make()
```

Section 3: Scoring and Analyzing Homology Models in Modeller and PyRosetta

The following scripts were used to create simple lists of scores in text files. These text files were then opened into an Excel spreadsheet that enabled models to be filtered and sorted by scores as desired. The first script is for validating models by N-DOPE Z scores: **zdope_scores.py**.

This script was run from the command line as follows:

[user@home]$ python zdope_scores.py file_list.txt

```
import sys
# This script computes a list of N-DOPE Z scores for several protein
structure files occupying the same folder.
# This script requires Modeller version 9.10 to run as well as a
simple text file with a file-list of all .pdb files to be assessed.

# Example for: model.assess_normalized_dope()
from modeller import *
frommodeller.scripts import complete_pdb

env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib')
env.libs.parameters.read(file='$(LIB)/par.lib')

# directories for input atom files
env.io.atom_files_directory = '.'

# Read a model previously generated by Modeller's automodel class
files = sys.argv[1]

f1 = open (files)
filename = ""
f2 = open ("zdope_scores.txt","a")

for line in f1:
     if(len(line)>1):
          filename = str.strip(line)
          mdl = complete_pdb(env, filename)
          zscore = mdl.assess_normalized_dope()
          #print str(zscore)
          f2.write(str(zscore)+" "+filename+"\n")
f2.close()

#Script originally written by Matthys Kroon modified by Benjamin
Kumwenda (22-08-2012).
#Script revised by Crystal-Leigh Clitheroe (12-09-2012).
```

Validating models by Rosetta Energy scores: **Renergy_scores.py**.

This script was run from the command line as follows:

[user@home]$ python Renergy_scores.py file_list.txt

```
import sys
# This script computes a list of Rosetta Energy scores for several
protein structure files occupying the same folder.
# This script requires Modeller version 9.10 and a
PyRosettainstalltion to run as well as a simple text file with a
file-list of all .pdb files to be assessed.

# Example for: standard score function(pose)

fromrosetta import *
rosetta.init()

scorefxn = create_score_function("standard")

# directories for input atom files
#Docenv.io.atom_files_directory = '.'

# Read a model previously generated by Modeller's automodel class

files = sys.argv[1]

f1 = open (files)
filename = ""
f2 = open ("Renergy_scores.txt","a")

for line in f1:
     if(len(line)>1):
          filename = str.strip(line)
          pose = pose_from_pdb(filename)
          Rscore = scorefxn(pose)
          #print str(Rscore)
          f2.write(str(Rscore)+" "+filename+"\n")
f2.close()

#Script modified from zdope_scores.py to run in the PyRosetta
environment by Crystal-Leigh Clitheroe (12-10-2012).
```

Validating models by Cα-RMSD scores: **rmsd_scores.py**.

This script was run from the command line as follows:

[user@home]$ python rmsd_scores.py file_list.txt

```
import sys
# This script computes a list of alpha-Carbon Root Mean Square
Deviation (Ca-RMSD) scores for several protein structure files
occupying the same folder.
# This script requires Modeller version 9.10 and a
PyRosettainstalltion to run as well as a simple text file with a
file-list of all .pdb files to be assessed. It also requires the
original template of the models built to be in the same folder.

# Example for: CA_rmsd(template_pose,model_pose)

fromrosetta import *
rosetta.init()

scorefxn = create_score_function("standard")

# Read a model previously generated by Modeller's automodel class

files = sys.argv[1]

template_pose = pose_from_pdb("template_file.pdb")

f1 = open (files)
filename = ""
f2 = open ("rmsd_scores.txt","a")

for line in f1:
     if(len(line)>1):
          filename = str.strip(line)
          model_pose = pose_from_pdb(filename)
          RMSDscore = CA_rmsd(template_pose, model_pose)
          #print str(RMSDscore)
          f2.write(str(RMSDscore)+" "+filename+"\n")
f2.close()

#Script modified from zdope_scores.py to run in the PyRosetta
environment by Crystal-Leigh Clitheroe (12-10-2012).
```

Section 4: Calculating Interaction and Binding Energies for Protein-protein Complexes

The following program was written to calculate approximate interaction and binding energies for the various protein-protein complexes studied in the project. The rationale and basic principles of the script are outlined in Chapter 4, Section 4.2.2. This script produces several files. It does this by splitting the complex file into separate chain entries that are written to their own .pdb files. Each of these chain files are scored for control and then minimized in PyRosetta in one of two ways, using either the classic_relax or repack_rotamers protocols. These modified files are then also scored and saved as separate entries. All the saved scores are then used to calculate the interaction and binding energies of the original complex. The rationale and basic principles of the script's calculations are outlined in Chapter 4, Section 4.2.2. The final output file is a text report. An example of the report file produced by this script follows the script details below. The current example of the binding energy calculations script is for the model 'HsTPR2ab2yeast_13_PTIEEVDcomplex.pdb: **int_bind_calc.py**

This script was run from the command line as follows:

[user@home]$ python int_bind_calc.py

```
# To assure the user that program is running:
print "Binding energy calculator for complexed proteins"

from Bio.PDB import *
import sys

#Import Rosetta libraries
fromrosetta import *
rosetta.init()
scorefxn = create_score_function("standard")     #set a global
scorefxn
relax = FastRelax()              #instantiates the refinement
protocol
relax.set_scorefxn(scorefxn)


#Import Modeller libraries
from modeller import *
frommodeller.scripts import complete_pdb
env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib')
env.libs.parameters.read(file='$(LIB)/par.lib')
# directories for input atom files
env.io.atom_files_directory = '.'


# Load the Complex
```

```
start = open('HsTPR2ab2yeast_13_PTIEEVDcomplex.pdb','r')
oldfile = start.readlines()

linelist = []

import re             # need to recognise text so import regex

forpatt in oldfile:
ifre.match('ATOM', patt):
linelist.append(patt)            # search for ATOM entries and split
                                 #file lines into field entries


chA = []
chB = []
chC = []
chD = []

chain = 0

for chain in linelist:
     if chain[21:22] == 'A':    #Split field entries into chain
info
          chA.append(chain)
     elif chain[21:22] == 'B':
          chB.append(chain)
     elif chain[21:22] == 'C':
          chC.append(chain)
     elif chain[21:22] == 'D':
          chD.append(chain)

# Create the new report file

report =
open("Energy_report_%s"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"),"w")
report.write("Binding energy report for
%s.pdb:\n\n"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"))


iflen(chA) != 0:
     newfileA =
open("%s_A.pdb"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"),"w")

     # Create the new .pdb files for chain A

     for j in chA:                          #For each line of the
original file:
          if j[:4]=="ATOM":                 #if an "ATOM" entry
               newfileA.write(j)            #copy line of old file
     newfileA.write('END')
     newfileA.close()

     chainfile_A = ("%s_A.pdb"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"))
```

```
      unmin_chA = pose_from_pdb(chainfile_A)
      Rosenergy_unminA = scorefxn(unmin_chA)
      nomA = complete_pdb(env, chainfile_A)
      zscore_unminA = nomA.assess_normalized_dope()

      rotam_chA = pose_from_pdb(chainfile_A)
      taskA = standard_packer_task(rotam_chA)
      taskA.restrict_to_repacking()
      mover = PackRotamersMover(scorefxn, taskA)
      mover.apply(rotam_chA)
      rotam_chA.dump_pdb("Rotamers_chain_A.pdb")

      Rosenergy_rotamA = scorefxn(rotam_chA)
      rotA = complete_pdb(env, "Rotamers_chain_A.pdb")
      zscore_rotamA = rotA.assess_normalized_dope()
      RMSD_rotA = CA_rmsd(unmin_chA, rotam_chA)

      relax_chA = pose_from_pdb(chainfile_A)
      relax.apply(relax_chA)
      relax_chA.dump_pdb("Relaxed_chain_A.pdb")

      Rosenergy_relaxA = scorefxn(relax_chA)
      relA = complete_pdb(env, "Relaxed_chain_A.pdb")
      zscore_relaxA = relA.assess_normalized_dope()
      RMSD_relA = CA_rmsd(unmin_chA, relax_chA)

      # Update the report file with Chain A info

      report.write("For Chain A:\n\nState\t\tDope-
Z\t\tRosetta\t\tRMSD\nUnminimised\t%.3f\t\t%.3f\t\tN/A\nRelaxed\t\t%
.3f\t\t%.3f\t\t%.3f\nRotamers\t%.3f\t\t%.3f\t\t%.3f\n\n"%(zscore_unm
inA, Rosenergy_unminA, zscore_relaxA, Rosenergy_relaxA, RMSD_relA,
zscore_rotamA, Rosenergy_rotamA, RMSD_rotA))


iflen(chB) != 0:
      newfileB =
open("%s_B.pdb"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"),"w")

      # Create the new .pdb files for chain B

      for k in chB:              #For each line of the original
file:
            if k[:4]=="ATOM":               #if an "ATOM" entry
                  newfileB.write(k)         #copy line of old file
      newfileB.write('END')
      newfileB.close()

      chainfile_B = ("%s_B.pdb"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"))

      unmin_chB = pose_from_pdb(chainfile_B)
      Rosenergy_unminB = scorefxn(unmin_chB)
      nomB = complete_pdb(env, chainfile_B)
      zscore_unminB = nomB.assess_normalized_dope()
```

```
      rotam_chB = pose_from_pdb(chainfile_B)
      taskB = standard_packer_task(rotam_chB)
      taskB.restrict_to_repacking()
      mover = PackRotamersMover(scorefxn, taskB)
      mover.apply(rotam_chB)
      rotam_chB.dump_pdb("Rotamers_chain_B.pdb")

      Rosenergy_rotamB = scorefxn(rotam_chB)
      rotB = complete_pdb(env, "Rotamers_chain_B.pdb")
      zscore_rotamB = rotB.assess_normalized_dope()
      RMSD_rotB = CA_rmsd(unmin_chB, rotam_chB)

      relax_chB = pose_from_pdb(chainfile_B)
      relax.apply(relax_chB)
      relax_chB.dump_pdb("Relaxed_chain_B.pdb")

      Rosenergy_relaxB = scorefxn(relax_chB)
      relB = complete_pdb(env, "Relaxed_chain_B.pdb")
      zscore_relaxB = relB.assess_normalized_dope()
      RMSD_relB = CA_rmsd(unmin_chB, relax_chB)

      # Update the report file with Chain B info

      report.write("For Chain B:\n\nState\t\tDope-
Z\t\tRosetta\tRMSD\nUnminimised\t%.3f\t\t%.3f\t\tN/A\nRelaxed\t\t%.3
f\t\t%.3f\t\t%.3f\nRotamers\t%.3f\t\t%.3f\t\t%.3f\n\n"%(zscore_unmin
B, Rosenergy_unminB, zscore_relaxB, Rosenergy_relaxB, RMSD_relB,
zscore_rotamB, Rosenergy_rotamB, RMSD_rotB))
else:
      Rosenergy_unminB = 0
      Rosenergy_relaxB = 0
      Rosenergy_rotamB = 0

iflen(chC) != 0:
      newfileC =
open("%s_C.pdb"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"),"w")

      # Create the new .pdb files for chain C

      for l in chC:              #For each line of the original
file:
            if l[:4]=="ATOM":               #if an "ATOM" entry
                  newfileC.write(l)         #copy line of old file
      newfileC.write('END')
      newfileC.close()

      chainfile_C = ("%s_C.pdb"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"))

      unmin_chC = pose_from_pdb(chainfile_C)
      Rosenergy_unminC = scorefxn(unmin_chC)
      nomC = complete_pdb(env, chainfile_C)
      zscore_unminC = nomC.assess_normalized_dope()

      rotam_chC = pose_from_pdb(chainfile_C)
      taskC = standard_packer_task(rotam_chC)
```

```
      taskC.restrict_to_repacking()
      mover = PackRotamersMover(scorefxn, taskC)
      mover.apply(rotam_chC)
      rotam_chC.dump_pdb("Rotamers_chain_C.pdb")

      Rosenergy_rotamC = scorefxn(rotam_chC)
      rotC = complete_pdb(env, "Rotamers_chain_C.pdb")
      zscore_rotamC = rotC.assess_normalized_dope()
      RMSD_rotC = CA_rmsd(unmin_chC, rotam_chC)

      relax_chC = pose_from_pdb(chainfile_C)
      relax.apply(relax_chC)
      relax_chC.dump_pdb("Relaxed_chain_C.pdb")

      Rosenergy_relaxC = scorefxn(relax_chC)
      relC = complete_pdb(env, "Relaxed_chain_C.pdb")
      zscore_relaxC = relC.assess_normalized_dope()
      RMSD_relC = CA_rmsd(unmin_chC, relax_chC)

      # Update the report file with Chain C info

      report.write("For Chain C:\n\nState\t\tDope-
Z\t\tRosetta\t\tRMSD\nUnminimised\t%.3f\t\t%.3f\t\tN/A\nRelaxed\t\t%
.3f\t\t%.3f\t\t%.3f\nRotamers\t%.3f\t\t%.3f\t\t%.3f\n\n"%(zscore_unm
inC, Rosenergy_unminC, zscore_relaxC, Rosenergy_relaxC, RMSD_relC,
zscore_rotamC, Rosenergy_rotamC, RMSD_rotC))

else:
      Rosenergy_unminC = 0
      Rosenergy_relaxC = 0
      Rosenergy_rotamC = 0



iflen(chD) != 0:
      newfileD =
open("%s_D.pdb"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"),"w")

      # Create the new .pdb files for chain D

      for m in chD:              #For each line of the original
file:
            if m[:4]=="ATOM":               #if an "ATOM" entry
                  newfileD.write(m)         #copy line of old file
      newfileD.write('END')
      newfileD.close()

      chainfile_D = ("%s_D.pdb"%("HsTPR2ab2yeast_13_PTIEEVDcomplex"))

      unmin_chD = pose_from_pdb(chainfile_D)
      Rosenergy_unminD = scorefxn(unmin_chD)
      nomD = complete_pdb(env, chainfile_D)
      zscore_unminD = nomC.assess_normalized_dope()

      rotam_chD = pose_from_pdb(chainfile_D)
```

```
      taskD = standard_packer_task(rotam_chD)
      taskD.restrict_to_repacking()
      mover = PackRotamersMover(scorefxn, taskD)
      mover.apply(rotam_chD)
      rotam_chD.dump_pdb("Rotamers_chain_D.pdb")

      Rosenergy_rotamD = scorefxn(rotam_chD)
      rotD = complete_pdb(env, "Rotamers_chain_D.pdb")
      zscore_rotamD = rotD.assess_normalized_dope()
      RMSD_rotD = CA_rmsd(unmin_chD, rotam_chD)

      relax_chD = pose_from_pdb(chainfile_D)
      relax.apply(relax_chD)
      relax_chD.dump_pdb("Relaxed_chain_D.pdb")

      Rosenergy_relaxD = scorefxn(relax_chD)
      relD = complete_pdb(env, "Relaxed_chain_D.pdb")
      zscore_relaxD = relD.assess_normalized_dope()
      RMSD_relD = CA_rmsd(unmin_chD, relax_chD)

      # Update the report file with Chain D info

      report.write("For Chain D:\n\nState\t\tDope-
Z\tRosetta\tRMSD\nUnminimised\t%.3f\t\t%.3f\t\tN/A\nRelaxed\t\t%.3f\
t\t%.3f\t\t%.3f\nRotamers\t%.3f\t\t%.3f\t\t%.3f\n\n"%(zscore_unminD,
Rosenergy_unminD, zscore_relaxD, Rosenergy_relaxD, RMSD_relD,
zscore_rotamD, Rosenergy_rotamD, RMSD_rotD))

else:
      Rosenergy_unminD = 0
      Rosenergy_relaxD = 0
      Rosenergy_rotamD = 0

# Score the overall complex

ori_complex = pose_from_pdb('HsTPR2ab2yeast_13_PTIEEVDcomplex.pdb')
Rosenergy_complex = scorefxn(ori_complex)

# Use unminimised versions of each chain to calculate the
interaction #Energy of the complex:

BE_unmin = Rosenergy_complex - Rosenergy_unminA - Rosenergy_unminB -
Rosenergy_unminC - Rosenergy_unminD

# Use two different types of minimisation (relaxed = minimised,
#rotamers = repack the rotamers) to calculate the binding energy of
#the complex

BE_relax = Rosenergy_complex - Rosenergy_relaxA - Rosenergy_relaxB -
Rosenergy_relaxC - Rosenergy_relaxD
BE_rotamers = Rosenergy_complex - Rosenergy_rotamA -
Rosenergy_rotamB - Rosenergy_rotamC - Rosenergy_rotamD

report.write("Comparative Binding
Energies:\n\nState\t\tEnergy\nUnminimised\t%.3f\nRelaxed\t\t%.3f\nRo
```

```
tamers\t%.3f\n\n"%(BE_unmin, BE_relax, BE_rotamers))

report.close()

#Script written by Crystal-Leigh Clitheroe (15-10-2012).
```

An example of the output file for this script:
Energy_report_HsTPR2ab2yeast_13_PTIEEVDcomplex.txt

```
Binding energy report for HsTPR2ab2yeast_13_PTIEEVDcomplex.pdb:

For Chain A:

State           Dope-Z          Rosetta         RMSD
Unminimised     -0.859          967.937         N/A
Relaxed         -1.041          -753.498        10.037
Rotamers        -0.913          4602.857        0.000

For Chain C:

State           Dope-Z     Rosetta         RMSD
Unminimised     0.218      6.338           N/A
Relaxed         0.477      -4.743          0.548
Rotamers        0.051      0.877           0.000

Comparative Binding Energies:

State           Energy
Unminimised     8.996
Relaxed         1741.512
Rotamers        -3620.463
```

Appendix 3: Phylogenetic Analysis

Section 1: Full-Protein Analysis

Bayesian Information Criterion (BIC) and Akaike Information Criterion (corrected, AIC^C) Analysis of evolutionary models:

Parameters: Complete site coverage, all others default.

Table A3.1: Top ten models selected from the lowest BIC and AIC^C Amino Acid Substitution Table scores in MEGA. The three yellow-highlighted models (best of each evolutionary model type) were used for analysis.

Model	#Parameters	BIC	AIC^c	lnL	Invariant	Gamma
tREV+G+I+F	138	58912	57754	-28738	0.044538	1.55811
rtREV+G+F	137	58929	57779	-28752	n/a	1.08544
WAG+G+I	119	59152	58153	-28957	0.046162	1.79716
WAG+G	118	59180	58189	-28976	n/a	1.23988
WAG+G+I+F	138	59282	58124	-28924	0.045915	1.70607
rtREV+G+I	119	59309	58310	-29035	0.042282	1.56793
WAG+G+F	137	59313	58163	-28944	n/a	1.16987
rtREV+G	118	59321	58331	-29047	n/a	1.12958
JTT+G+I+F	138	59521	58363	-29043	0.045948	1.52534
JTT+G+F	137	59546	58396	-29060	n/a	1.05874

First Phylogenetic ML Tree compared to its Bootstrap consensus Tree:

Using rtREV +F +G +I Model, Gamma value of 2, 500 Bootstrapped replicates, complete site coverage (all other parameters default).

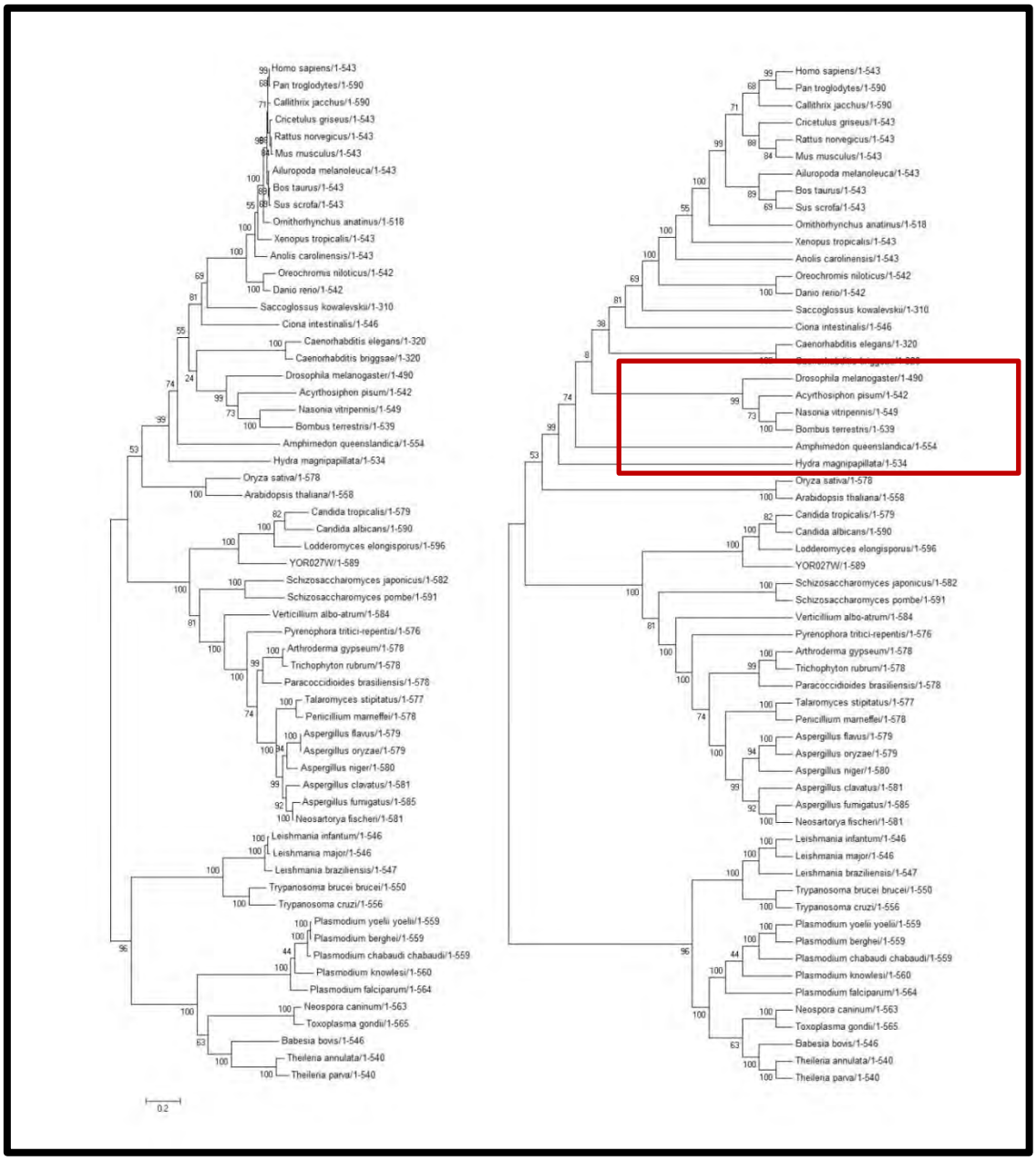


Figure A3.1: The Maximum Likelihood (ML) tree using the Reverse Transcriptase (rtREV) evolutionary model (left) shows good branching and topological agreement with its bootstrap consensus tree (right).There is one area of minor topological disagreement,This shows that the rtREV model is a good model to use for the Hop protein tree.

Second Phylogenetic ML Tree compared to its Bootstrap consensus Tree:

Using WAG +F +G, Gamma value of 2, 500 Bootstrapped replicates, complete site coverage (all other parameters default).

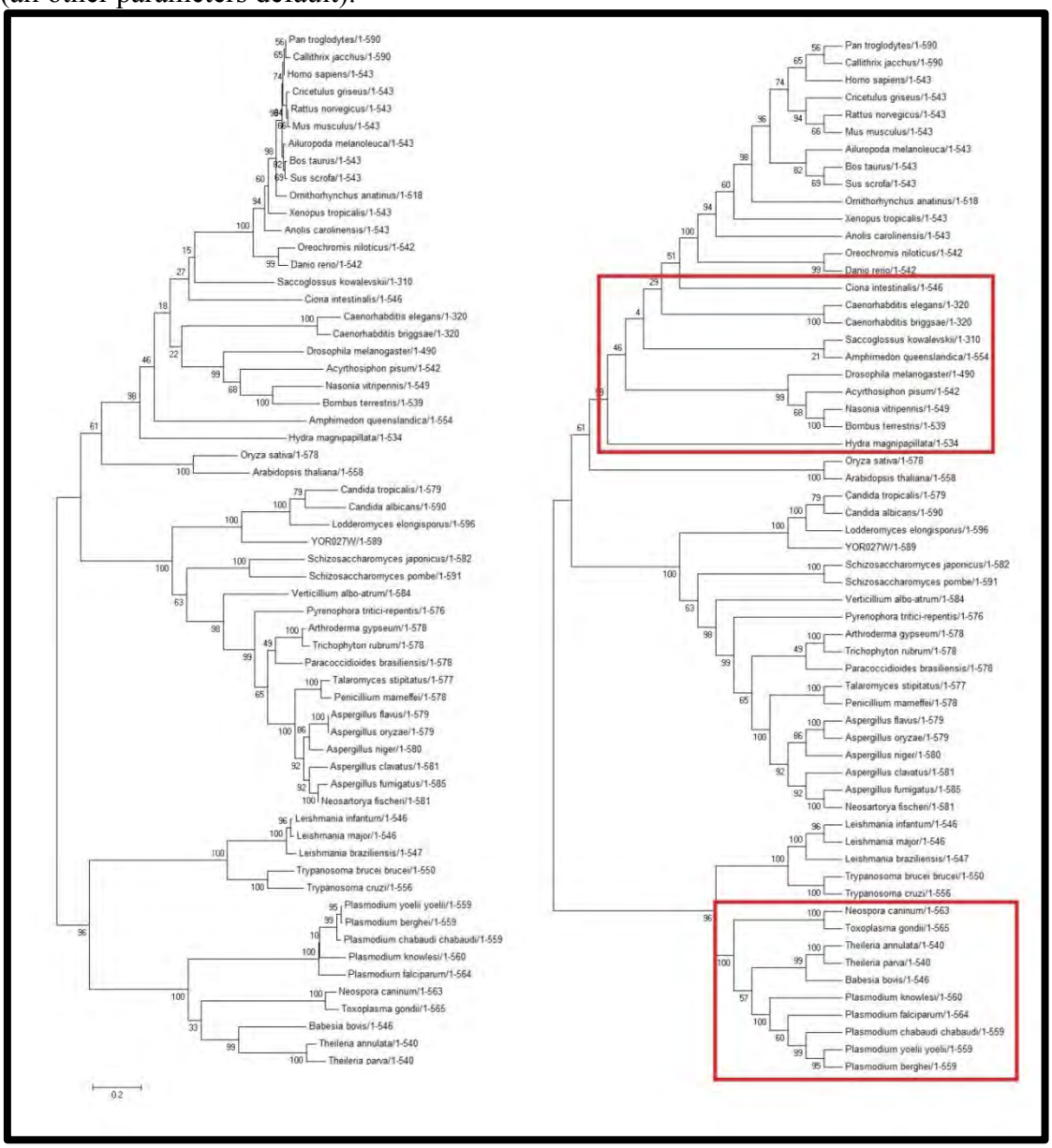


Figure A3.2: A third ML tree was built using the Whelan and Goldman (WAG) model, with complete site coverage (left). There are two areas of major topological disagreement, in the invertebrates and the Apicomplexa, this indicates a poor evolutionary model.

Invertebrates: In the ML tree (left), *S. kowalevski* is the first outgroup to the vertebrates, whereas the consensus tree places *C. intestinalis* as the first outgroup to the vertebrates. In the ML tree the next outgroup is an ingroup of the nematodes and the insects (followed by *A. queenslandica* as the next outgroup), whereas in the consensus tree the nematodes are the next

outgroup to *C. intestinalis* and vertebrates, with an ingroup consisting of *S. kowalevski* and *A. queenslandica* being the next outgroup to the nematodes and all others (followed by the insects). *H. magnipapilata* is the final outgroup for all vertebrates and invertebrates in both trees.

Apicomplexa: In the ML tree, *T. gondii* and *N. caninum* group are outgrouped to the rest of the non-euglenazoanapicomplexans, while in the consensus tree, the Plasmodium genus is the outgroup to the rest of the Apicomplexa. Within the plasmodial genus, *P. falciparum* is outgrouped to the rest of the species, while the consensus tree groups *P. knowlesi* outside of the rest of the species.

Third Phylogenetic ML Tree compared to its Bootstrap consensus Tree:

Using JTT +F +G +I, Gamma distributed rates2, 500 Bootstrapped replicates, complete site coverage (all other parameters default).

Figure A3.3: A second ML tree was built using the Jones, Taylor and Thornton (JTT) model, with complete site coverage (left). There is only minor topological disagreement;*A. gypseum, T. rubrum* and *P. brasiliensis* are a single out-group to innermost fungal group on the bootstrap consensus tree (right), whereas *A. gypseum* and *T. rubrum* are the outgroup to innermost group plus *P. brasiliensis* on the ML tree.

Comparing the Evolutionary models

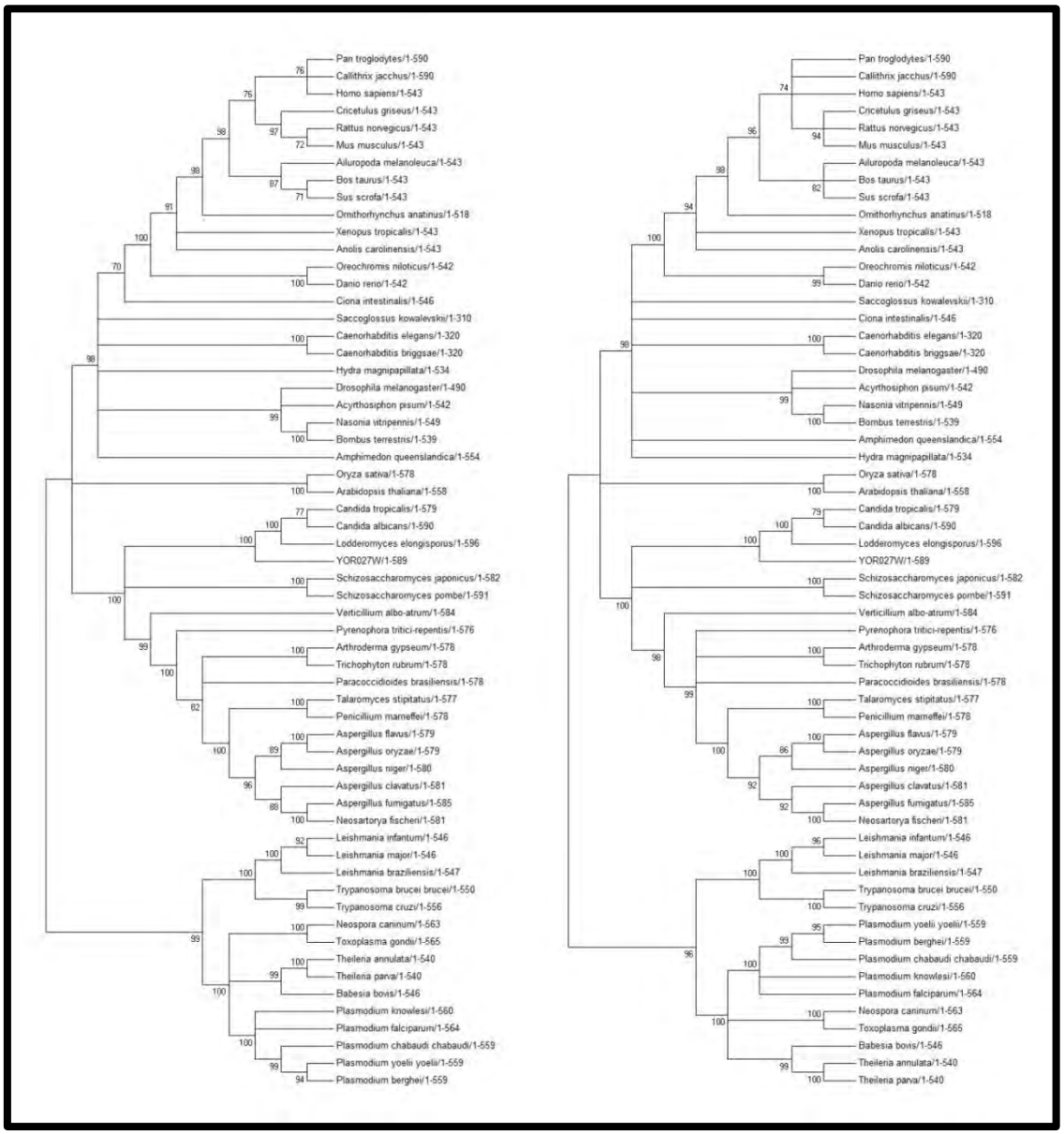


Figure A3.4: Comparing the condensed trees (at 70% cutoff) produced from two different evolutionary models, JTT (Left) and WAG (right). There is generally greater loss of branch resolution (e.g. within the mammals, fungi and invertebrates) in the tree produced with the WAG model.

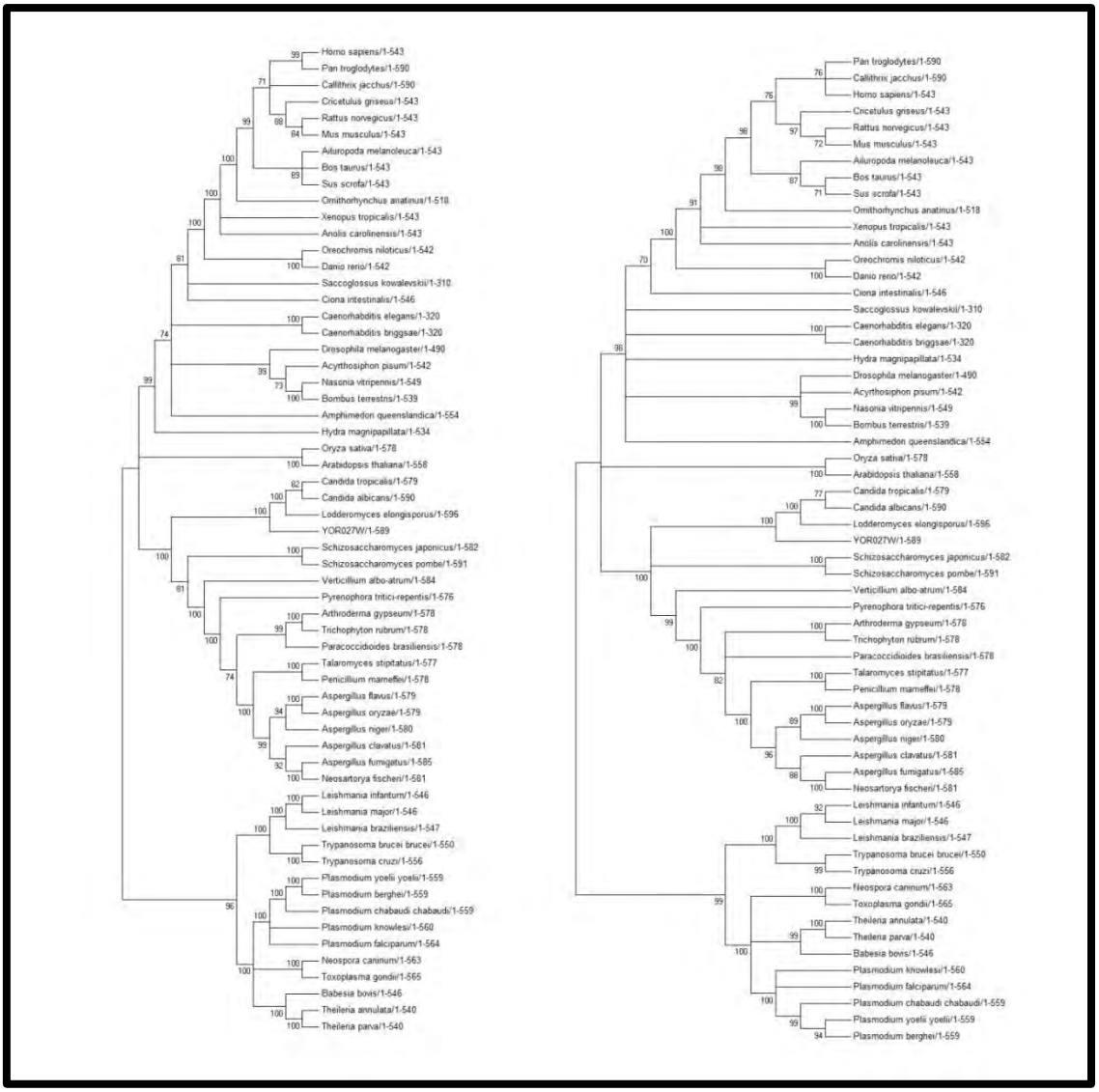


Figure A3.5: Comparing the condensed trees (at 70% cutoff) produced from two different evolutionary models, JTT (right) and rtREV (left). There is generally some loss of branch resolution (e.g. within the mammals, fungi and invertebrates) in the tree produced with the JTT model.

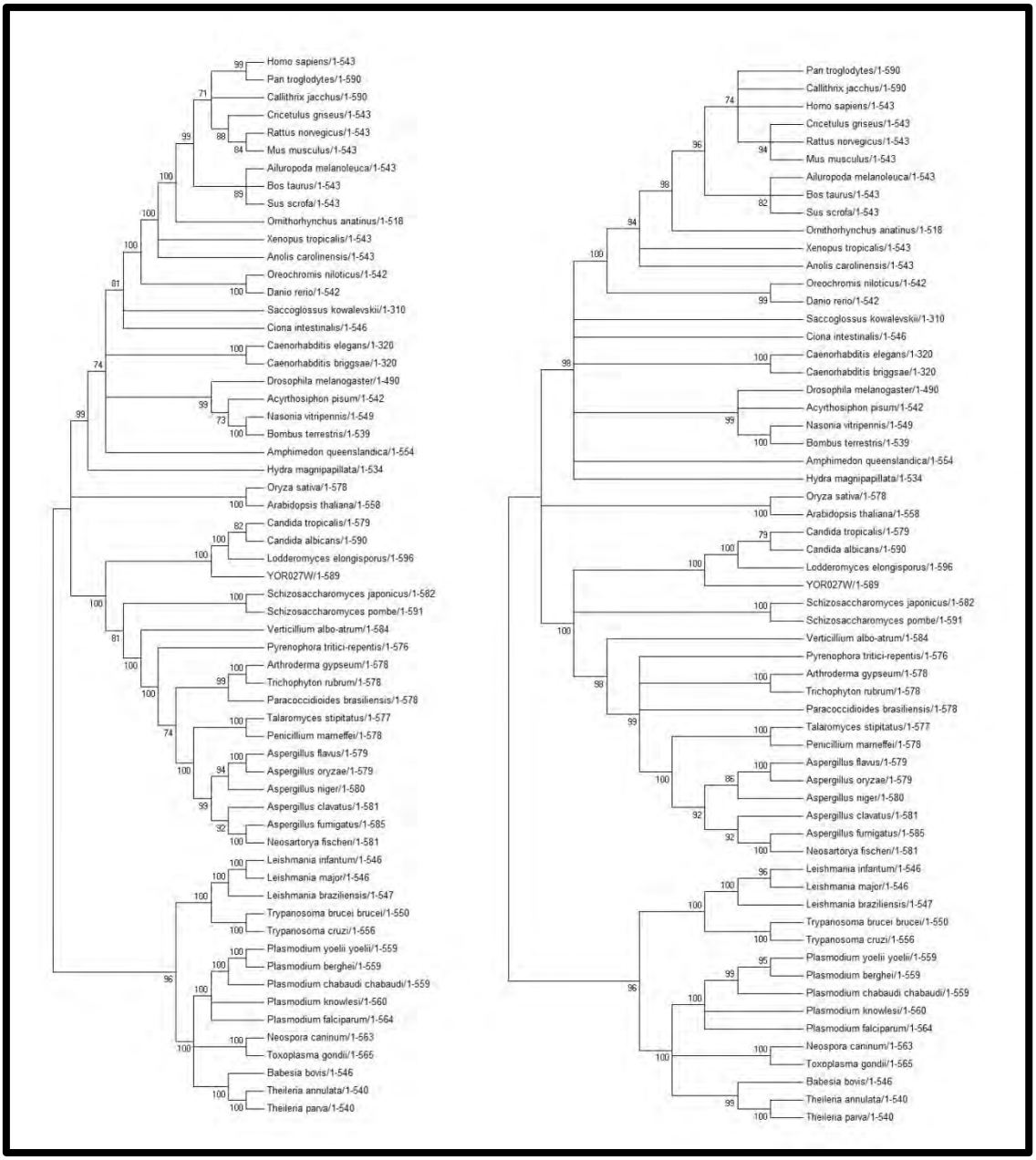


Figure A3.6: Comparing the condensed trees (at 70% cutoff) produced from two different evolutionary models, WAG (right) and rtREV (left). There is generally some loss of branch resolution (e.g. within the mammals, fungi and invertebrates) in the tree produced with the WAG model.

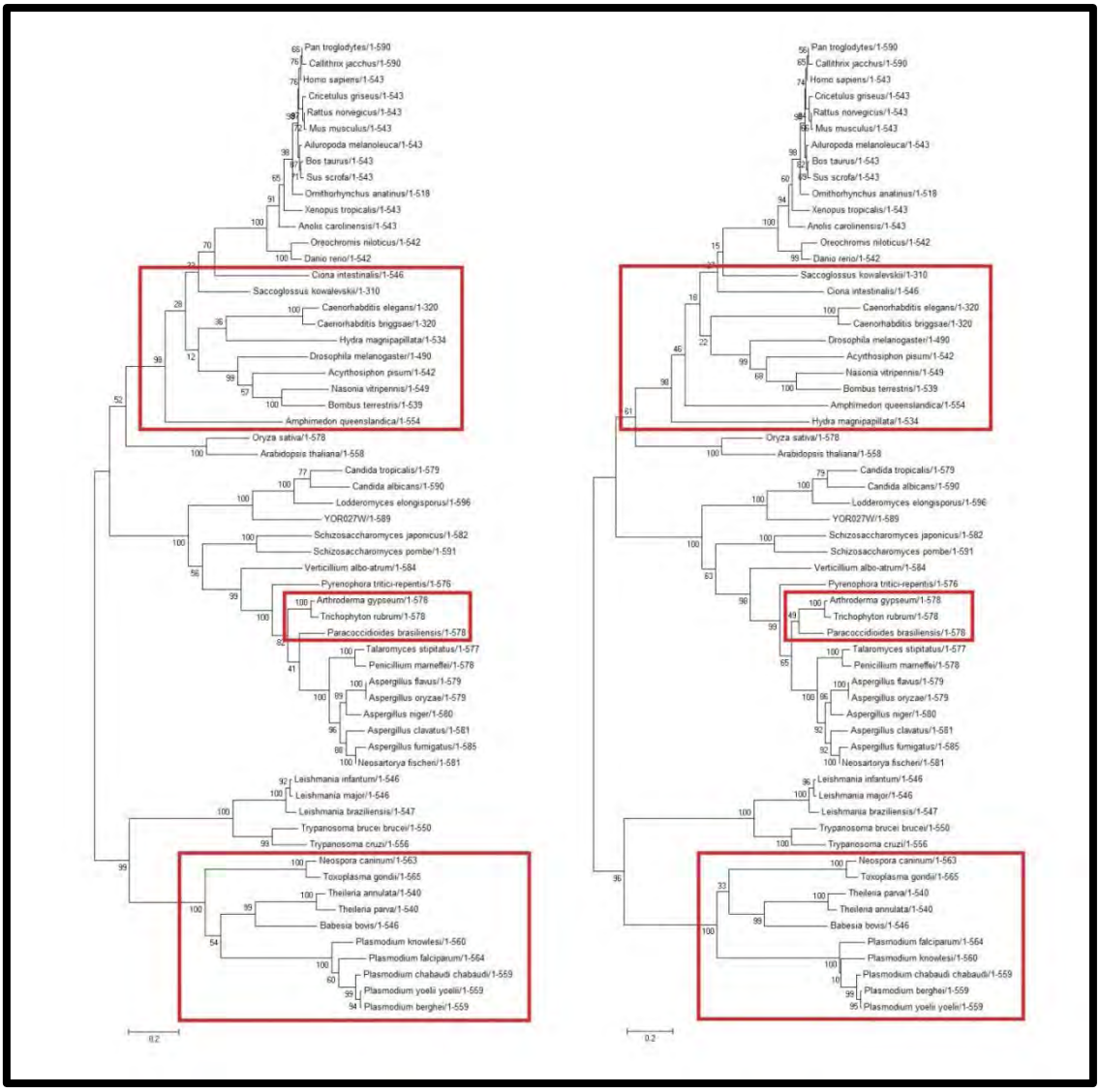


Figure A3.7: Comparing Maximum Likelihood trees produced with the JTT Model (left) and with the WAG model (right). There are three areas of topological disagreement, in the invertebrates, the fungi and the non-euglenazoanapicomplexans.

Invertebrates: In the WAG tree (right), *S. kowalevski* is the first outgroup to the vertebrates (with *C. intestinalis*following), whereas in the JTT tree the situation is reversed. In the JTT tree, the next outgroup is an ingroup of the nematodes, *H. magnipapillata* and the insects (followed by *A. queenslandica* as the final outgroup to the vertebrates and invertebrates), whereas in the WAG tree*H. magnipapillata* is the final outgroup for all vertebrates and invertebrates.

Fungi: *A. gypseum, T. rubrum* and *P. brasiliensis* are a single out-group to innermost fungal group on the WAG tree (right), whereas *A. gypseum* and *T. rubrum* are the outgroup to innermost group plus *P. brasiliensis* on the JTT tree.

Apicomplexa: In the JTT tree, the*T. gondii* and *N. caninum*group are outgroup to the

remainingapicomplexans, while in the WAG tree, places the *Plasmodium* genus as the outgroup. Within the WAG plasmodialspecies sub-tree, *P. falciparum* is outgroup to the rest of the species, while in the JTT plasmodialspecies sub-tree,*P. knowlesi*is the outgroup. It is interesting to note that the WAG bootstrap consensus sub-tree (figure B) for this region is similar to that of JTT tree (figure C).

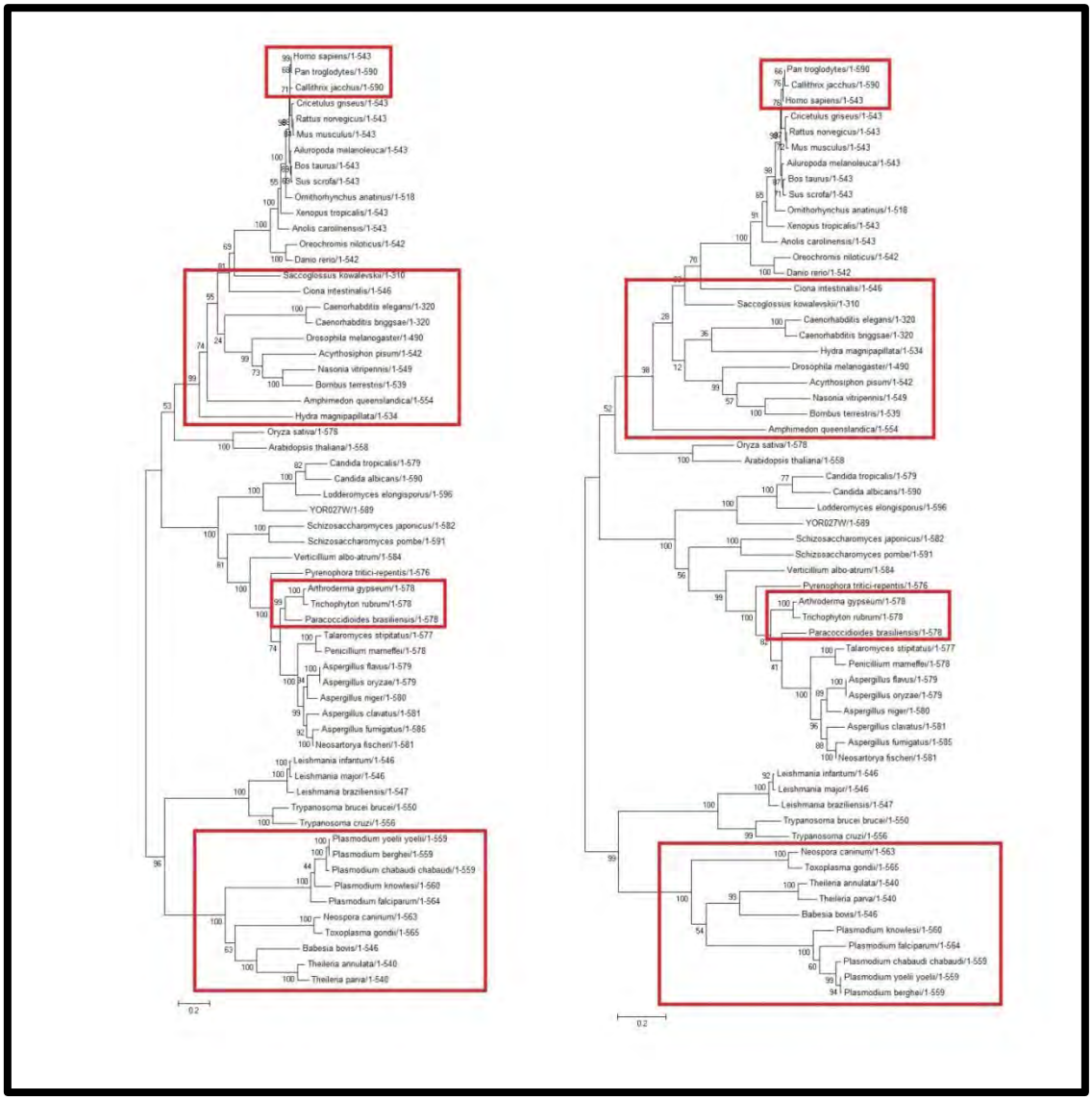


Figure A3.8: Comparing Maximum Liklihood trees produced with the JTT Model (right) and with the rtREVmodel (left). There are four areas of topological disagreement, in the primates, the invertebrates, the fungi and the non-euglenazoanapicomplexans.

Primates: In the JTT tree *H. sapiens* is an outgroup to *P. troglodytes* and *C. jacchus*, whereas in the rtREV tree *C. jacchus* is the outgroup to *P. troglodytes* and *H.sapiens*.

Invertebrates: In the rtREVtree ,*S. kowalevski* is the first outgroup to the vertebrates (with *C. intestinalis*following), whereas in the JTT tree the situation is reversed. In the JTT tree, the

next outgroup is an ingroup of the nematodes, *H. magnipapillata* and the insects (followed by *A. queenslandica* as the final outgroup to the vertebrates and invertebrates), whereas in the rtREV tree *H. magnipapillata* is the final outgroup for all vertebrates and invertebrates.

Fungi: *A. gypseum, T. rubrum* and *P. brasiliensis* are a single out-group to innermost fungal group on the rtREV tree, whereas *A. gypseum* and *T. rubrum* are the outgroup to innermost group including *P. brasiliensis* on the JTT tree (right).

Apicomplexa: In the JTT tree, the *T. gondii* and *N. caninum* group are outgroup to the remaining apicomplexans, while in the rtREV tree, places the *Plasmodium* genus as the outgroup. Within the rtREVplasmodial species sub-tree, *P. falciparum* is outgroup to the rest of the species, while in the JTT plasmodial species sub-tree, *P. knowlesi* is the outgroup.

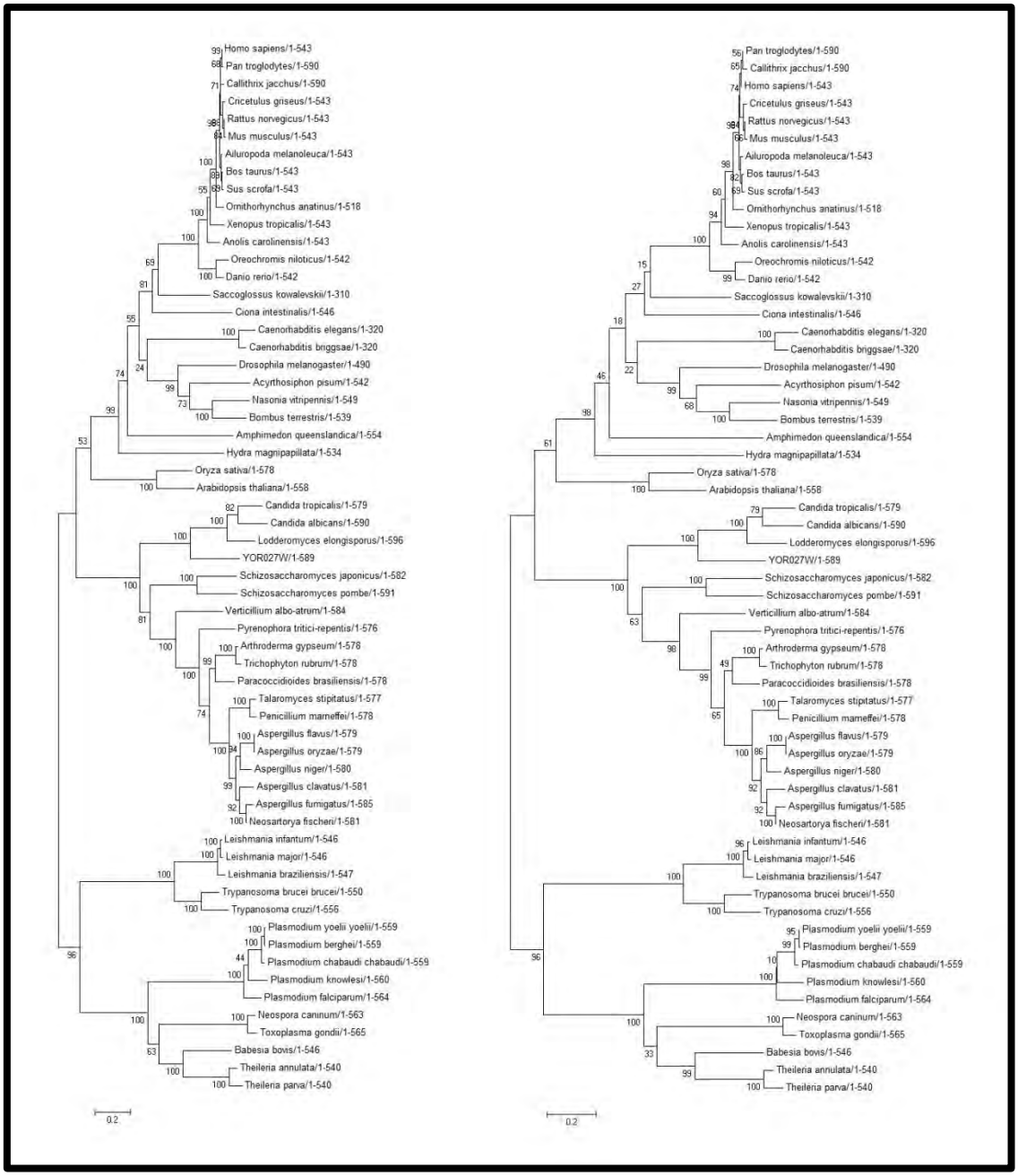


Figure A3.9: Comparing Maximum Likelihood trees produced with the rtREV Model (left) and with the WAG model (right). There is one area of topological disagreement; In the WAG tree *H. sapiens* is an outgroup to *P. troglodytes* and *C. jacchus*, whereas in the rtREV tree *C. jacchus* is the outgroup to *P. troglodytes* and *H. sapiens*. However, the rtREV tree appears to have greater bootstrap support values (in most but not all regions) and shorter branch lengths.

This, in combination with the good agreement with its bootstrapped tree, seems to indicate that thertREV model returns the most correct Maximum Likelihood protein tree for Hop.

A closer look at the JTT Model tree:

Ungapped Phylogenetic JTT ML Tree compared to its Bootstrap consensus Tree:

Using JTT +F +G +I, Gamma value of 2, 500 Bootstrapped replicates, 85% site coverage and all other parameters default.

Figure A3.10: Minor topological disagreement In the ML (left) tree *H. sapiens* is an outgroup to *C. Jacchus*and *P. troglodytes*, whereas in its bootstrapped tree (right), *C. jacchus*is an outgroup to *H. sapiens* and *P. troglodytes*.

Comparing the full coverage and 85% coverage ML trees:

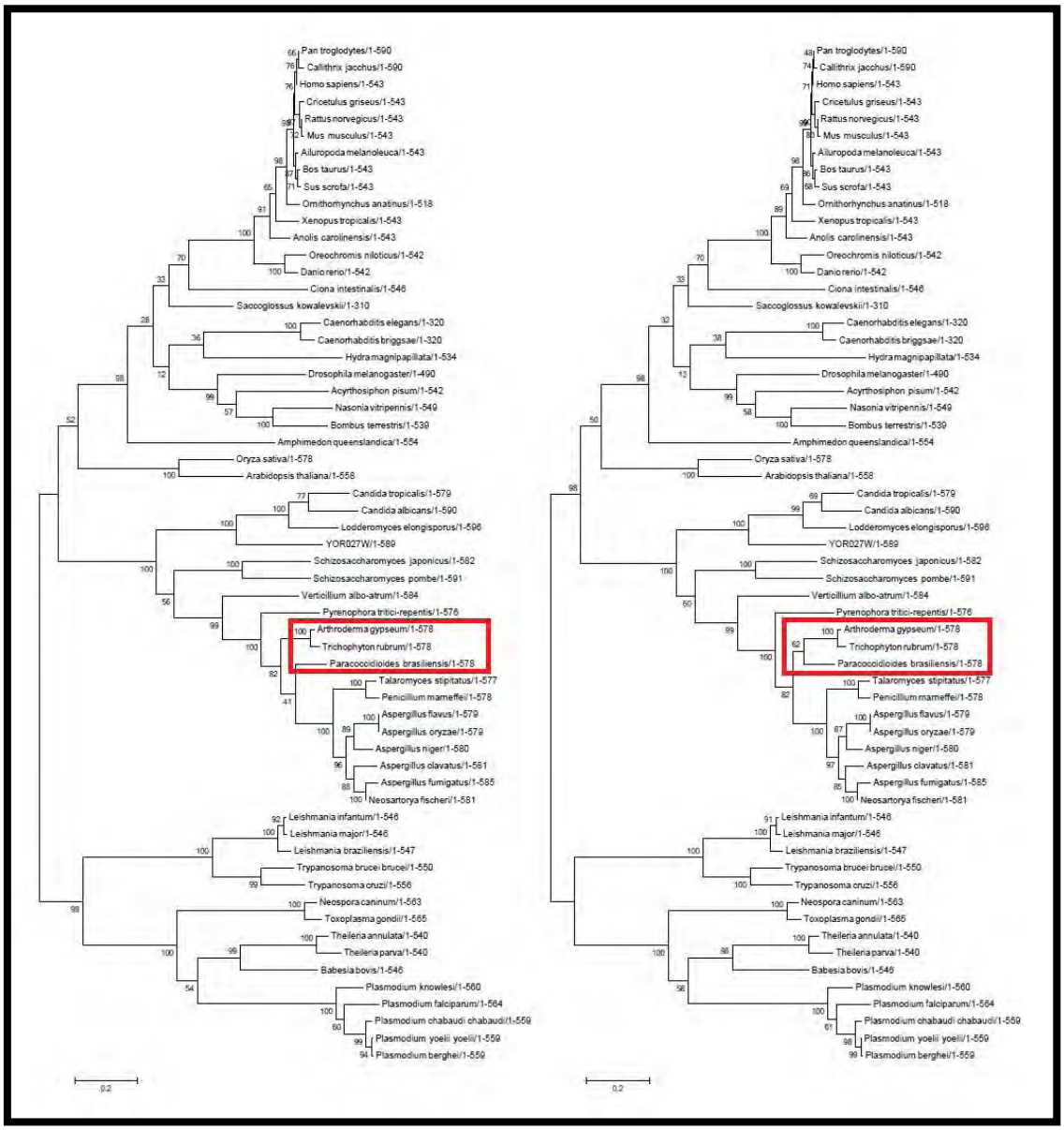


Figure A3.11: The same disparity in topological agreement as in figure A is observed: *A. gypseum*, *T. rubrum* and P. brasiliensis are a single out-group to innermost fungal group on the Bootstrap consensus tree, whereas *A. gypseum* and *T. rubrum* are the outgroup to innermost group plus *P. brasiliensis* on the ML Tree.

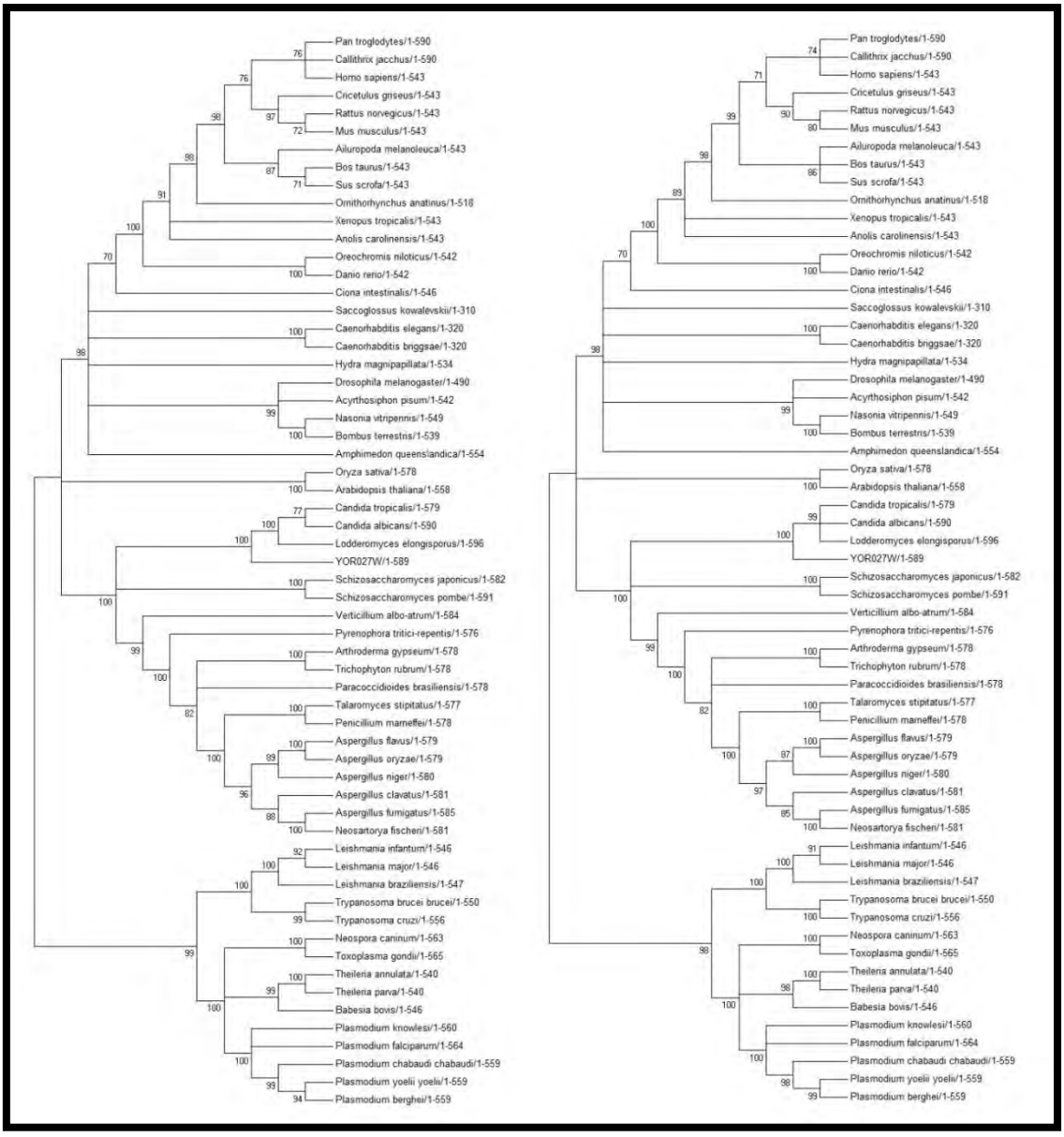


Figure A3.12: A look at the condensed versions of the above trees (cut off at confidence scores of 70), shows that the 85% coverage tree loses slightly more resolution of the branching (e.g. in the mammals and fungi), but overall, the trees are almost identical.

First Phylogenetic Neighbor-Joining Tree compared to its Bootstrap consensus Tree:

Using JTT +F +G +I Model, Gamma value of 2, 500 Bootstrapped replicates, complete site coverage (all other parameters default)

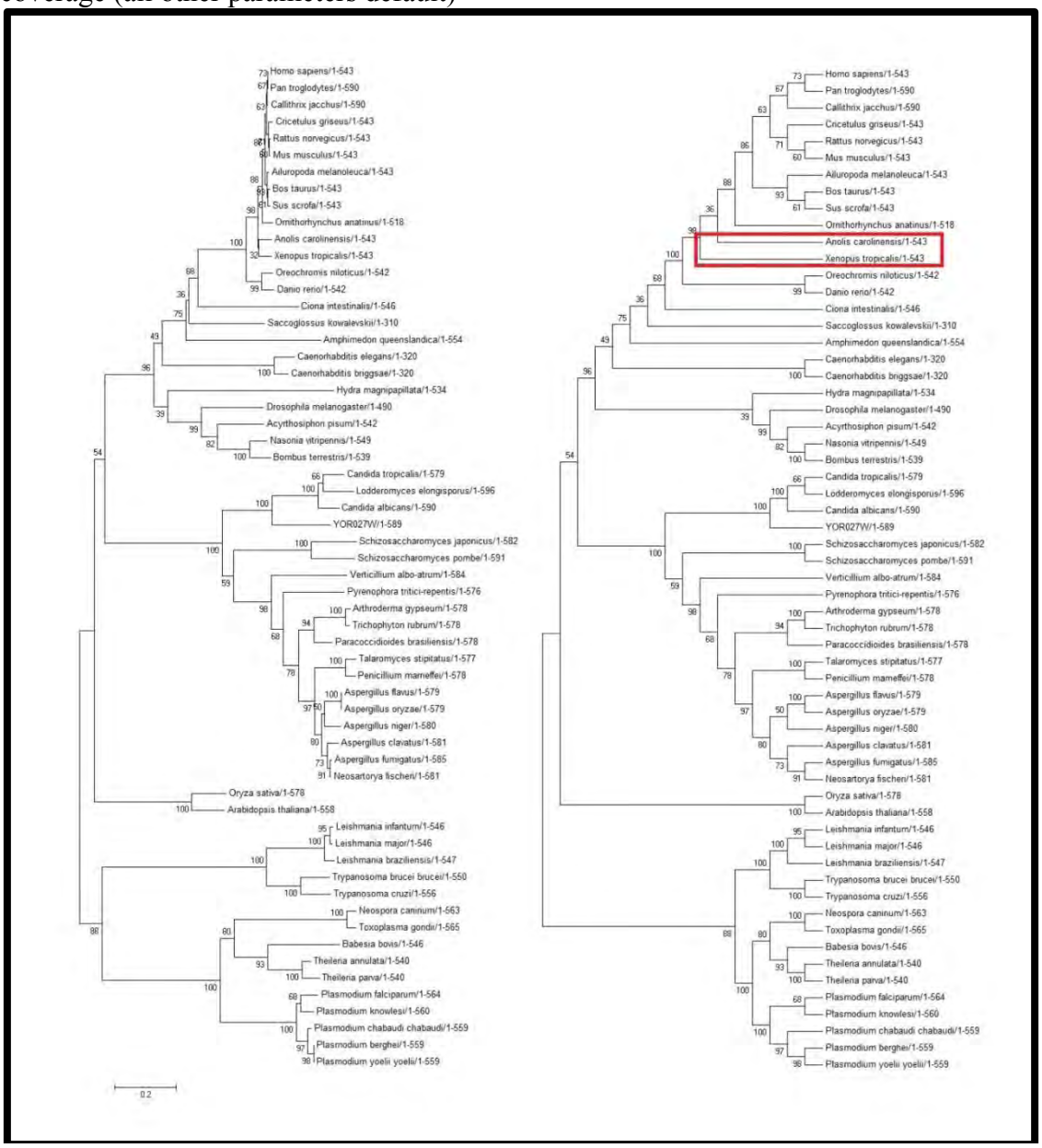


Figure A3.13: The Neighbor-Joining (NJ) tree using the JTT evolutionary model (left) shows good branching and topological agreement with its bootstrap consensus tree (right), with only one region of disagreement. In the NJ tree, *X. tropicalis* and *N. carolensis* are grouped, whereas in the bootstrap consensus tree, they are separate outgroups. This shows that the JTT model is a good model to use for the Hop Neighbour-Joining protein tree.

Comparing Tree-building Methods

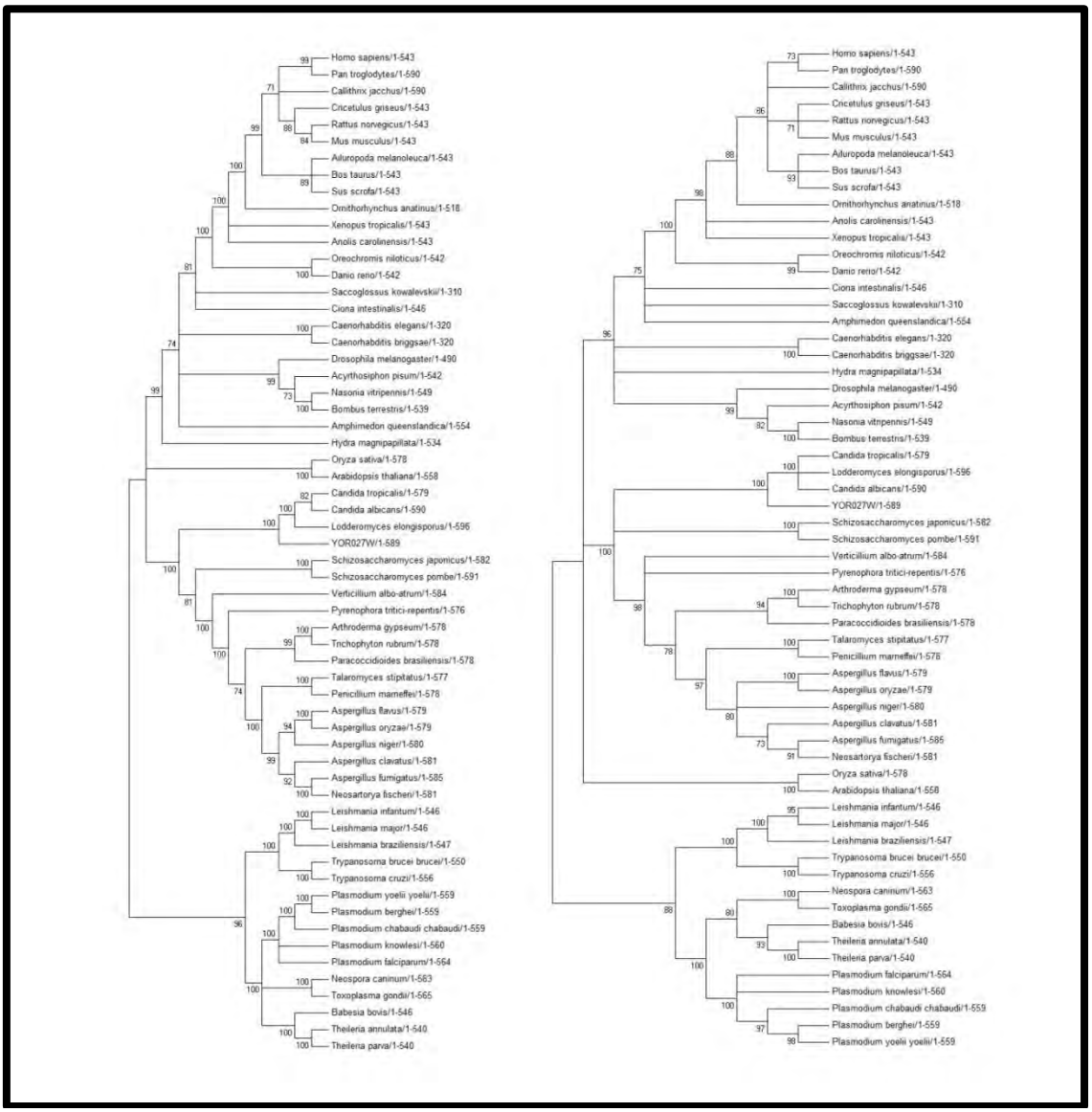
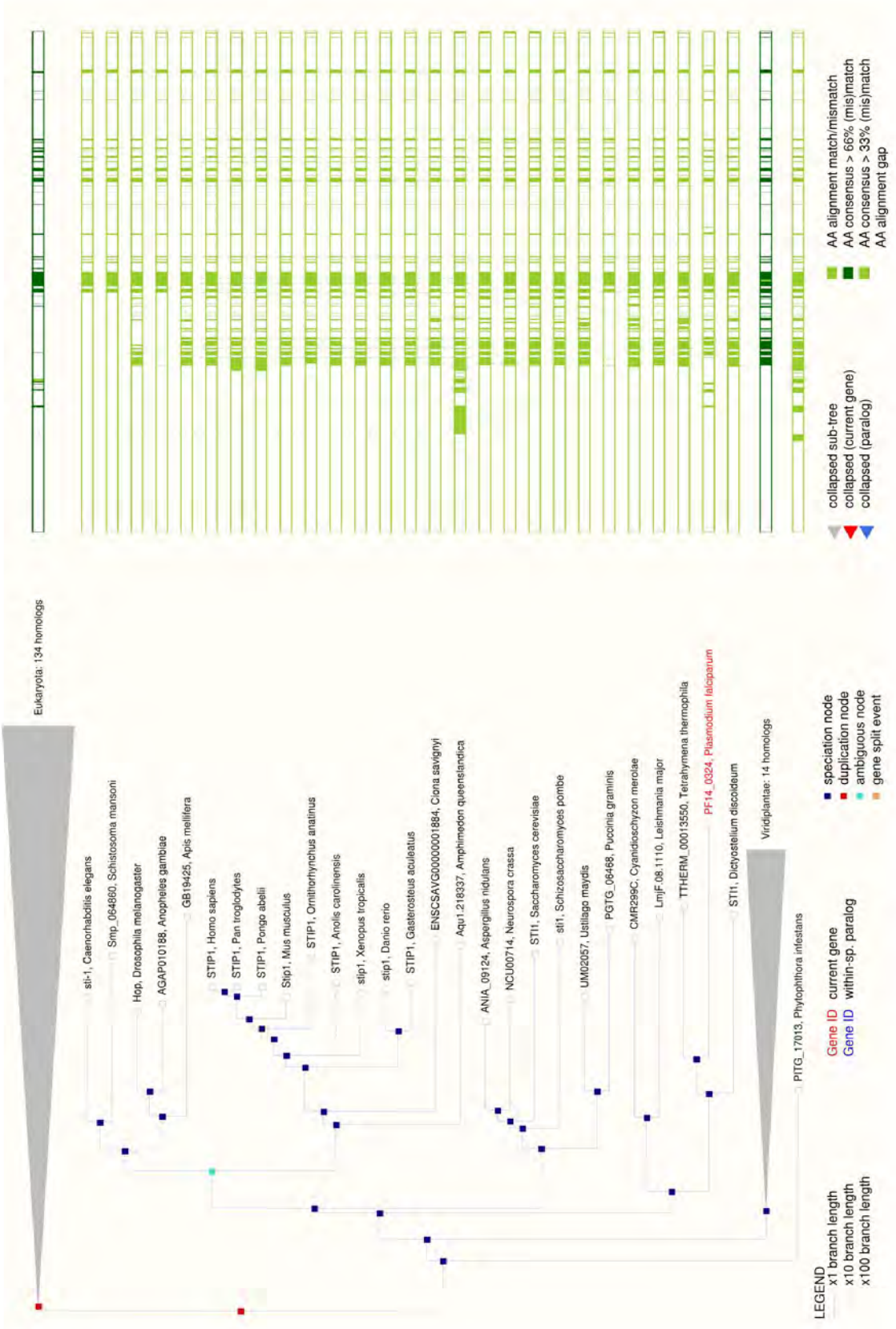


Figure A3.14: Comparison of the condensed (70%) versions of the rtREV ML tree (left) against the NJ tree (right); shows that the NJ tree loses more resolution of the branching (e.g. in the mammals and fungi), but overall, the trees are similar.

Appendix 4: Major Taxonomic Groups in the Compara-Gene Tree for PfHop on EnsEMBL



Appendix 5: Meme Results

Section 1: Single Motif Search

Meme Parameters: Full-length Protein, width 2-150, 1 motif search, multiple occurrences per sequence, all other parameters default.

Figure A5.1: Motif 1 for single motif search

Regex:
Y[YF][QN]K[SA]L[TV]EH[RN][TN][PR][DE][TV][LR][TKN]KLR[EN][AL]E[KR]AK[KE][KE][AE]E[RK]EAYI[DN]PE[KLE]AE[EK][AEH][RK]E[KL]GN[EKQ][KYF]F[KQ][EK][GA][DK][FWY]P[GE]A[VK]K[AHE]Y[TD]E[AM][IT][KR]R[NA]P[DN]D[AP][KR][GL][YF]SNRAA[AC][LY][TI]KL[ML][AE][FY]P[LSQ]A[LV][KQ]DC[DE][KE][AC]I[EK][LR]DP[KT]F[IV][KR][AG]YIRK

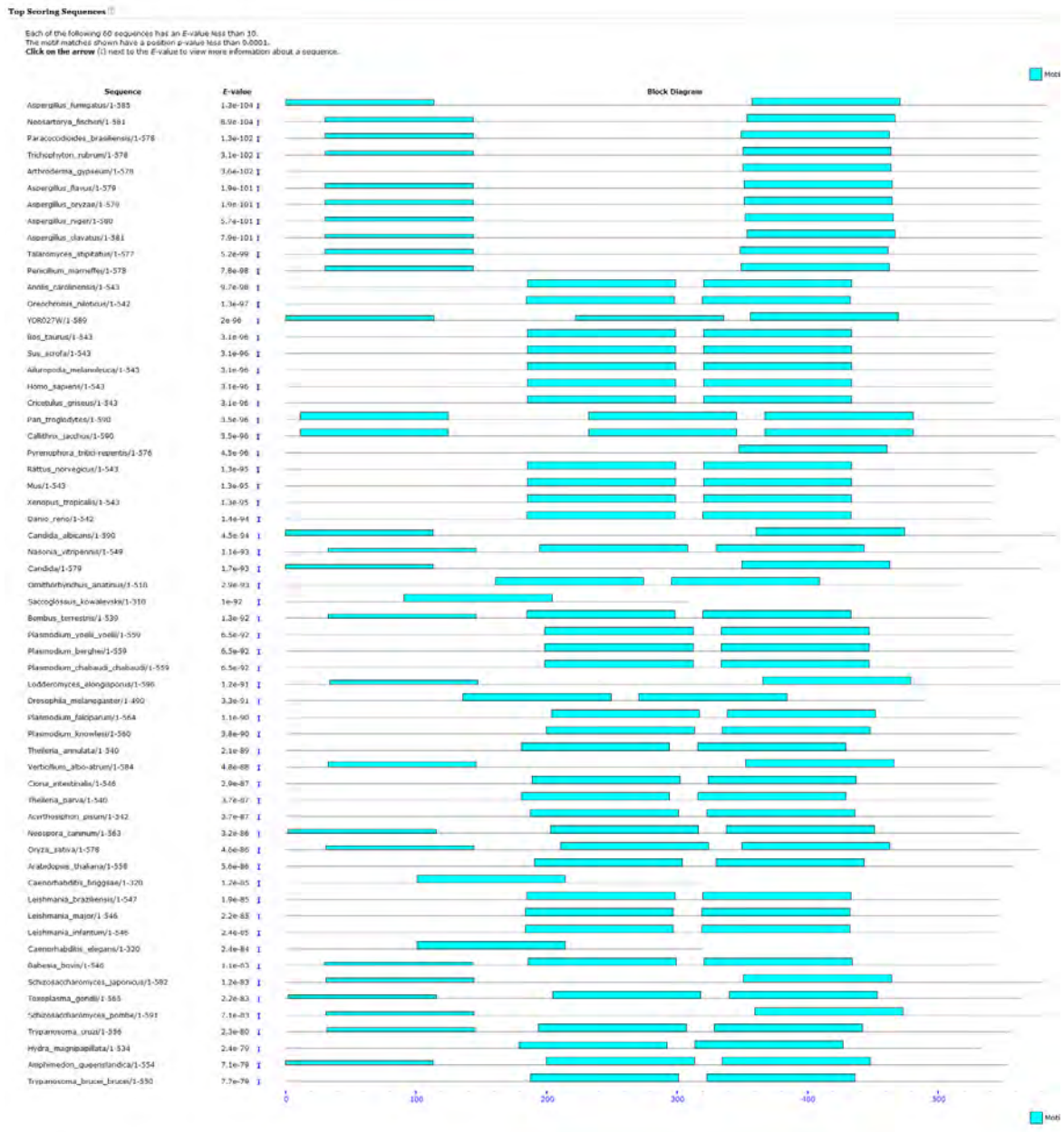


Figure A5.2: MAST results for all sequences, for single motif search

Section 2: Five Motif Search

Meme Parameters: Full-length Protein, motif width 2-150, 5 motif search, multiple occurrences per sequence, all other parameters default.

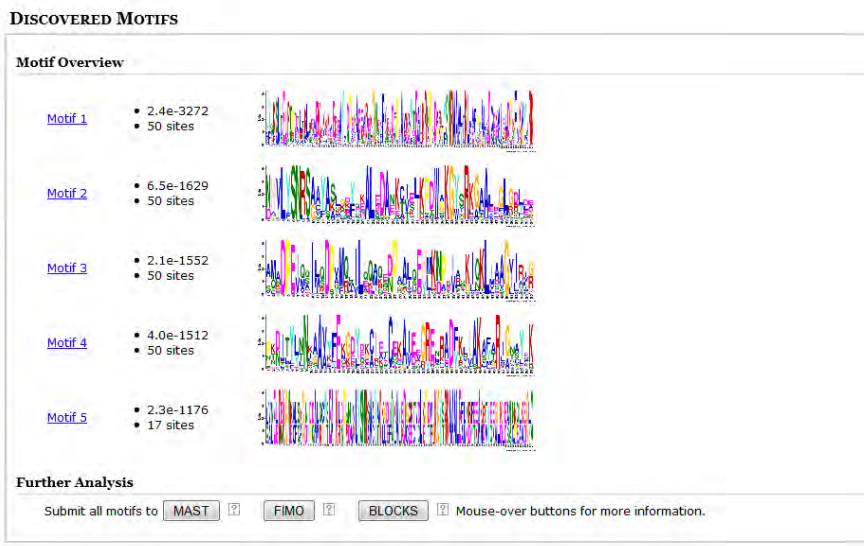


Figure A5.3: All motifs and scores

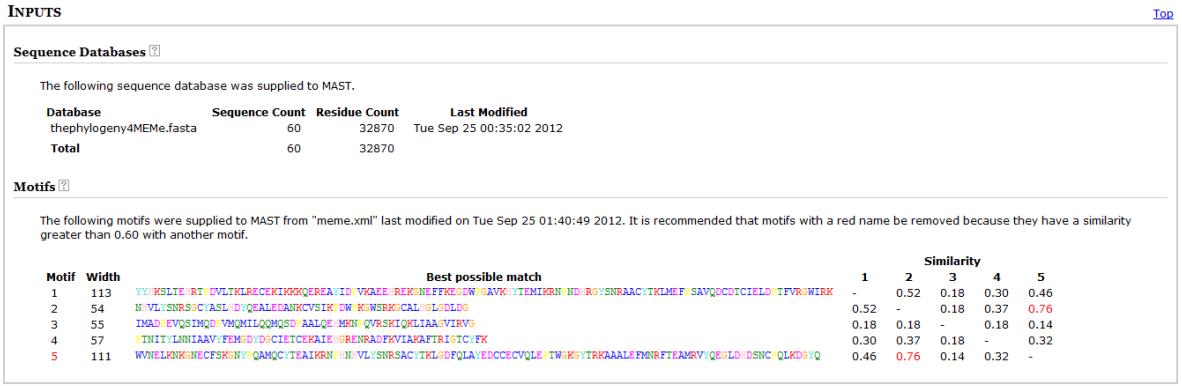


Figure A5.4: Mast analysis for 5 motif search

Motif Results:

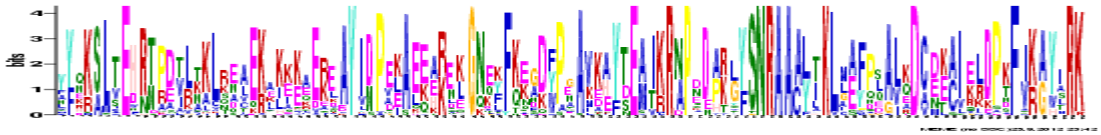


Figure A5.5: Motif 1 for 5 motif search

Regex:

Y[YF][QN]K[SA]L[TV]EH[RN][TN][PR][DE][TV][LR][TKN]KLR[EN][AL]E[KR]AK[KE][KE][AE]E[RK]EAYI[DN]PE[KLE]AE[EK][AEH][RK]E[KL]GN[EKQ][KYF]F[KQ][EK][GA][DK][FWY]P[GE]A[VK]K[AHE]Y[TD]E[AM][IT][KR]R[NA]P[DN]D[AP][KR][GL][YF]SNRAA[AC][LY][TI]KL[ML][AE][FY]P[LSQ]A[LV][KQ]DC[DE][KE][AC]I[EK][LR]DP[KT]F[IV][KR][AG]YIRK

Figure A5.6: Motif 2 for 5 motif search

Regex:
[ND]H[VI]LYSNRS[AG]AYA[SAK]L[GK][DK][YF]Q[KE]ALEDA[NE]K[CT][IVT][ES][IL]KPDW[APG]KG[YW]SRK[GA]AA[LE]HGL[GR][DR][LY][DEL]E

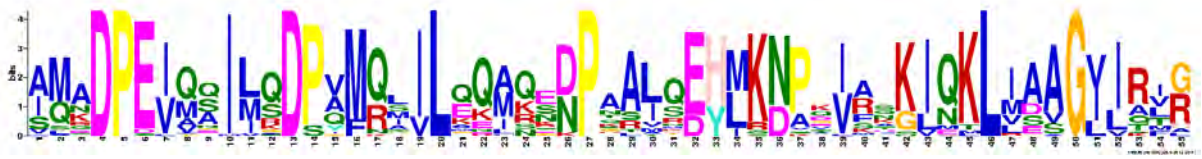


Figure A5.7: Motif 3 for 5 motif search

Regex:
[ND]H[VI]LYSNRS[AG]AYA[SAK]L[GK][DK][YF]Q[KE]ALEDA[NE]K[CT][IVT][ES][IL]KPDW[APG]KG[YW]SRK[GA]AA[LE]HGL[GR][DR][LY][DEL]E

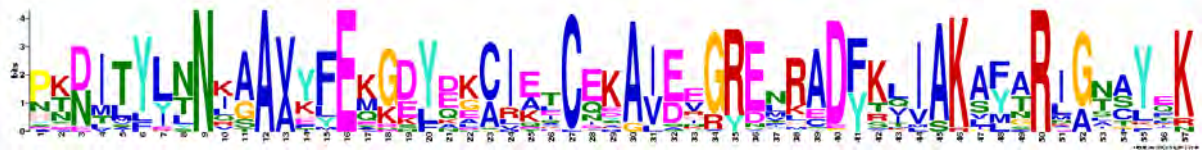


Figure A5.8: Motif 4 for 5 motif search

Regex:
P[KT][DN]ITYL[NT]N[KI][AG]A[VA][YK]FE[KM][GK][DE]Y[DEQ][KG][CA]I[EA]TC[EQ]KA[IV][ED][EV]GRENRAD[FY]KLIAK[AS][FY][AT]R[IL][GA][NT][ASC]Y[EQ]K

Figure A5.9: Motif 5 for 5 motif search

Regex:
[LQ][VA][NL]E[LE]K[EN]KGN[KE][AC][LF][SQ][KA]G[ND][IY][DP][DQ]A[LM][QK][CH]Y[ST]EAIK[LR][DN]P[KQ][ND][HA][KV]LYSNR[SA]A[AC]Y[AT]K[KL][GL][DE][YF]Q[KL]A[YL][EK]D[GC][CE][KE][TC][VI][DQ]L[KE]P[DT][WF][GI]KGY[ST]RKAAALE[FA][LM][NK][RD][FY][ET][EK]A[KM][DR][TV]Y[EQ][EK][GA]L[KD][HL][ED][AS][NS][NC][KP][QE][LA][KA][ED]G[LY]Q

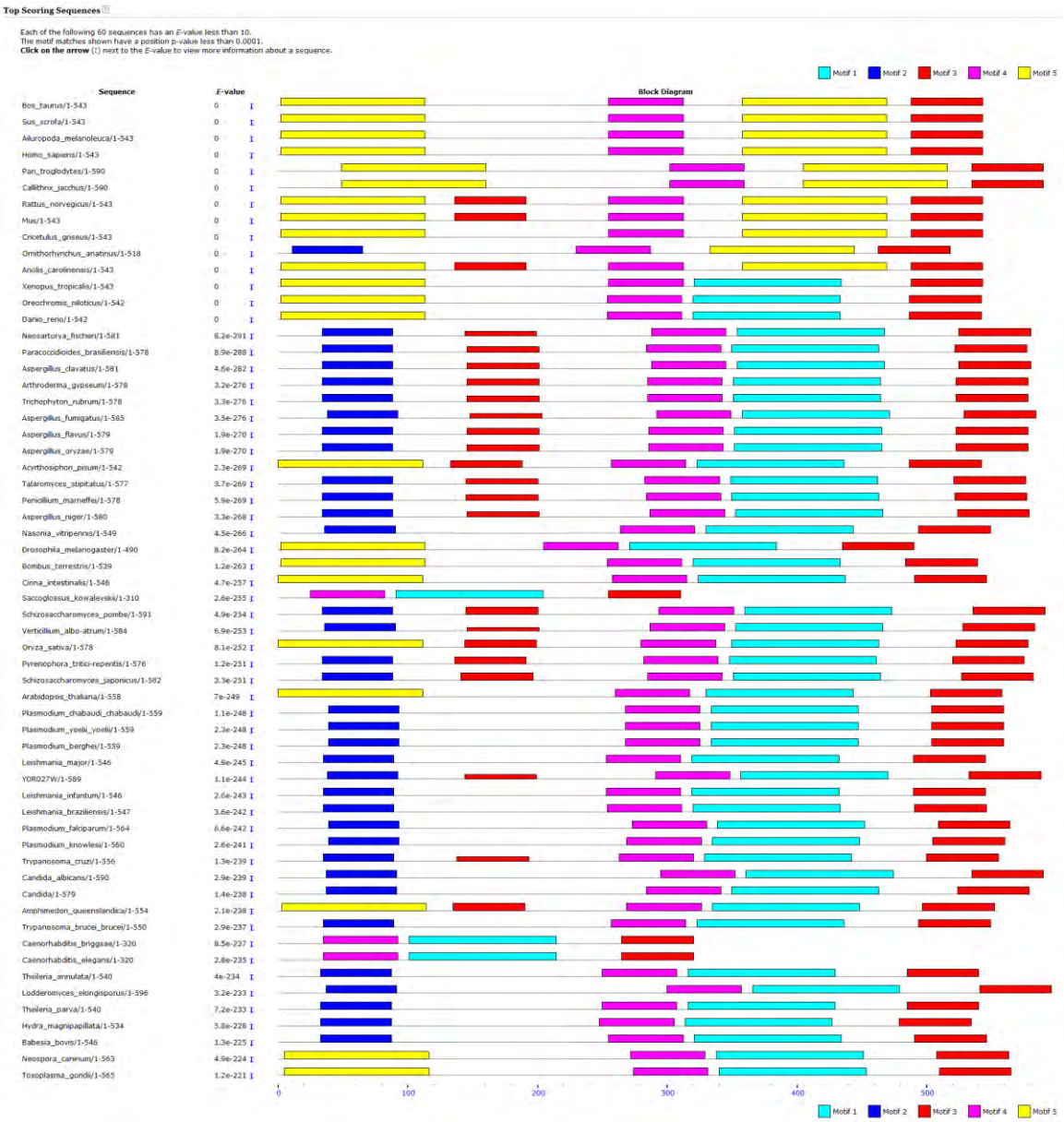


Figure A5.10: Mast results for all sequences submitted to Meme for 5 motif search

Section 3: Multiple Motif Search

Meme Parameters: Full-length Protein, motif width 2-150, 5, 10 and 20 motif searches, multiple occurrences per sequence, all other parameters default.

Figure A5.11: All motifs and scores from a 10 (top left), 15 (top right) and 20 (bottom) motif search.

From Figure A5.11, it is obvious that the maximum amount of motifs that Meme finds is 11, so only one motif set (from the 20 motif search) was analysed further.

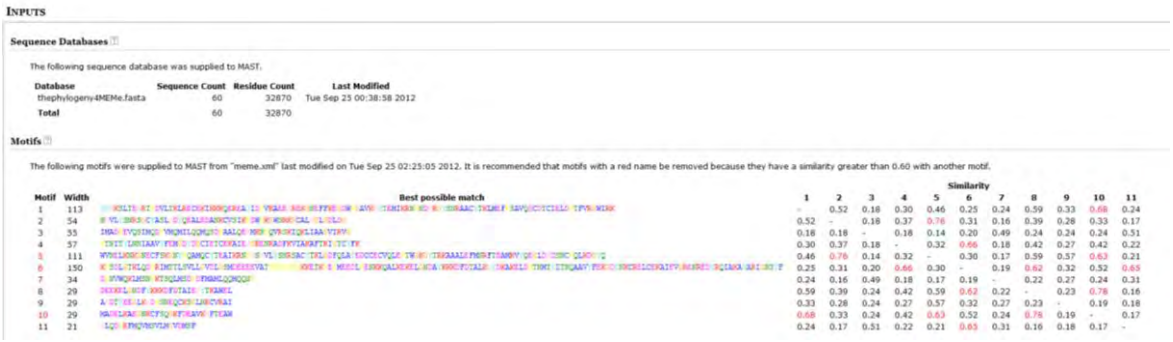


Figure A5.12: Mast analysis for 20 motif search

Motif Logos:

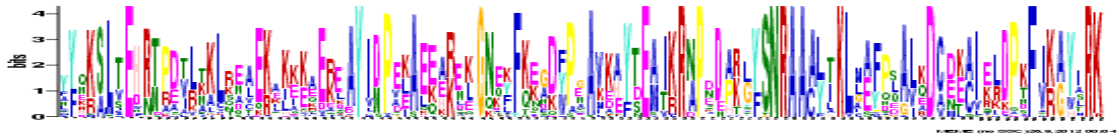


Figure A5.13: Motif 1 for 20 motif search

Regex:
Y[YF][QN]K[SA]L[TV]EH[RN][TN][PR][DE][TV][LR][TKN]KLR[EN][AL]E[KR]AK[KE][KE][AE]E[RK]EAYI[DN]PE[KLE]AE[EK][AEH][RK]E[KL]GN[EKQ][KYF]F[KQ][EK][GA][DK][FWY]P[GE]A[VK]K[AHE]Y[TD]E[AM][IT][KR]R[NA]P[DN]D[AP][KR][GL][YF]SNRAA[AC][LY][TI]KL[ML][AE][FY]P[LSQ]A[LV][KQ]DC[DE][KE][AC]I[EK][LR]DP[KT]F[IV][KR][AG]YIRK

Figure A5.14: Motif 2 for 20 motif search

Regex:
[ND]H[VI]LYSNRS[AG]AYA[SAK]L[GK][DK][YF]Q[KE]ALEDA[NE]K[CT][IVT][ES][IL]KPDW[APG]KG[YW]SRK[GA]AA[LE]HGL[GR][DR][LY][DEL]E

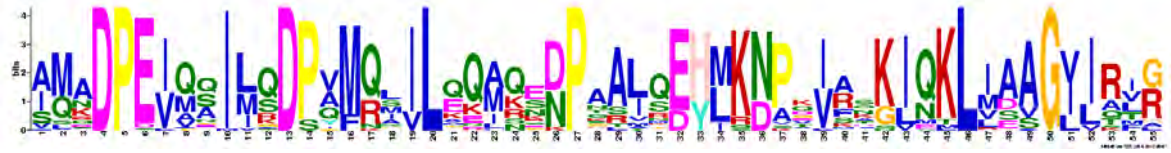


Figure A5.15: Motif 3 for 20 motif search

Regex:
[AI][MQ]ADPE[IV][QM][QSA]I[LM][QS]DP[VAQ]M[QR]LIL[QE]Q[AM][QK][ES][DN]PAA[LI][QS][ED][HY][ML]K[ND]PK[IV][AR]x[KG]I[QN]KL[IM]AAG[IVL][IL]R[ILV][GR]

Figure A5.16: Motif 4 for 20 motif search

Regex:
P[KT][DN]ITYL[NT]N[KI][AG]A[VA][YK]FE[KM][GK][DE]Y[DEQ][KG][CA]I[EA]TC[EQ]KA[IV][ED][EV]GRENRAD[FY]KLIAK[AS][FY][AT]R[IL][GA][NT][ASC]Y[EQ]K

Figure A5.17: Motif 5 for 20 motif search

Regex:
[LQ][VA][NL]E[LE]K[EN]KGN[KE][AC][LF][SQ][KA]G[ND][IY][DP][DQ]A[LM][QK][CH]Y[ST]EAIK[LR][DN]P[KQ][ND][HA][KV]LYSNR[SA]A[AC]Y[AT]K[KL][GL][DE][YF]Q[KL]A[YL][EK]D[GC][CE][KE][TC][VI][DQ]L[KE]P[DT][WF][GI]KGY[ST]RKAAALE[FA][LM][NK][RD][FY][ET][EK]A[KM][DR][TV]Y[EQ][EK][GA]L[KD][HL][ED][AS][NS][NC][KP][QE][LA][KA][ED]G[LY]Q

Figure A5.18: Motif 6 for 20 motif search

Regex:
KPSDLGTKLQDPR[IV]MTTLSVLLGVDLGSMDEEEE[VAI]ATPPPPPPPKKE[TP]KPEPMEEDLPENKKQALKEKELGN[DE]AYKKKDFD[TK]ALKHYD[KR]AK[ED]LDPTNMTY[IM]TNQAAV[YH]FEKGDY[NG]KCRELCEKAIEVGRENREDYRQIAKAYARIGNSYF

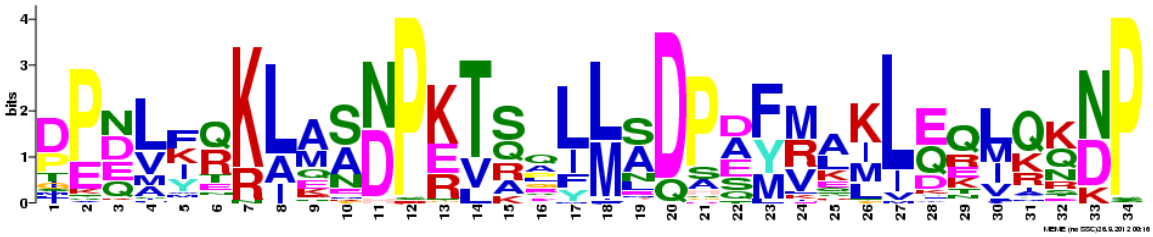


Figure A5.19: Motif 7 for 20 motif search

Regex:
[DP][PE][DNE][LV][FIK][QR][KR][LA][AM][SA][DN]P[KER][TV][SQ]Q[LI][LM][SA]DP[ADE][FYM][MRV][AL][KI]L[EQ]Q[LIM][QK][KQ][ND]P

Figure A5.20: Motif 8 for 20 motif search

Regex:
D[AKE][EL]K[EK][LK]GN[AED][AF][YF]K[KA]K[DQ]F[DE][ET]A[IL]EH[YF]TKA[IW]EL

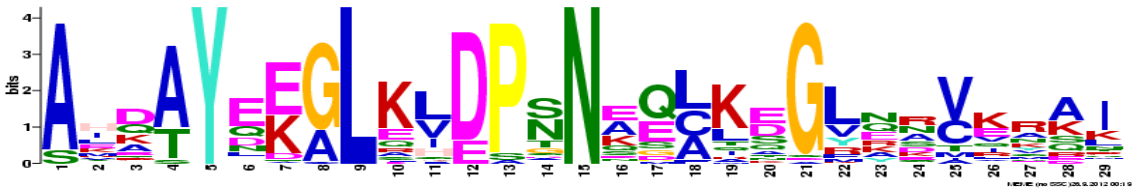


Figure A5.21: Motif 9 for 20 motif search

Regex:
AHD[AT]YE[EK][GA]LK[LV]DP[NS]N[EA][QE][LCA]K[ED]GLNR[VC]KR[AK]I

Figure A5.22: Motif 10 for 20 motif search

Regex:
[EM]AD[EA]LK[AE][EK]GN[KE][AL][FY][KS][AQ][KG][KD][FY][ED]EA[ILV][KE]K[YF][TS][EKQ]AI

Figure A5.23: Motif 11 for 20 motif search

Regex:
[EH][LM][KQ][DN]P[RV][FIV][LMA][QT][VTK][LMI][SGQ][VK]L[LM][GD][VI][DG][LM][SI][FA]

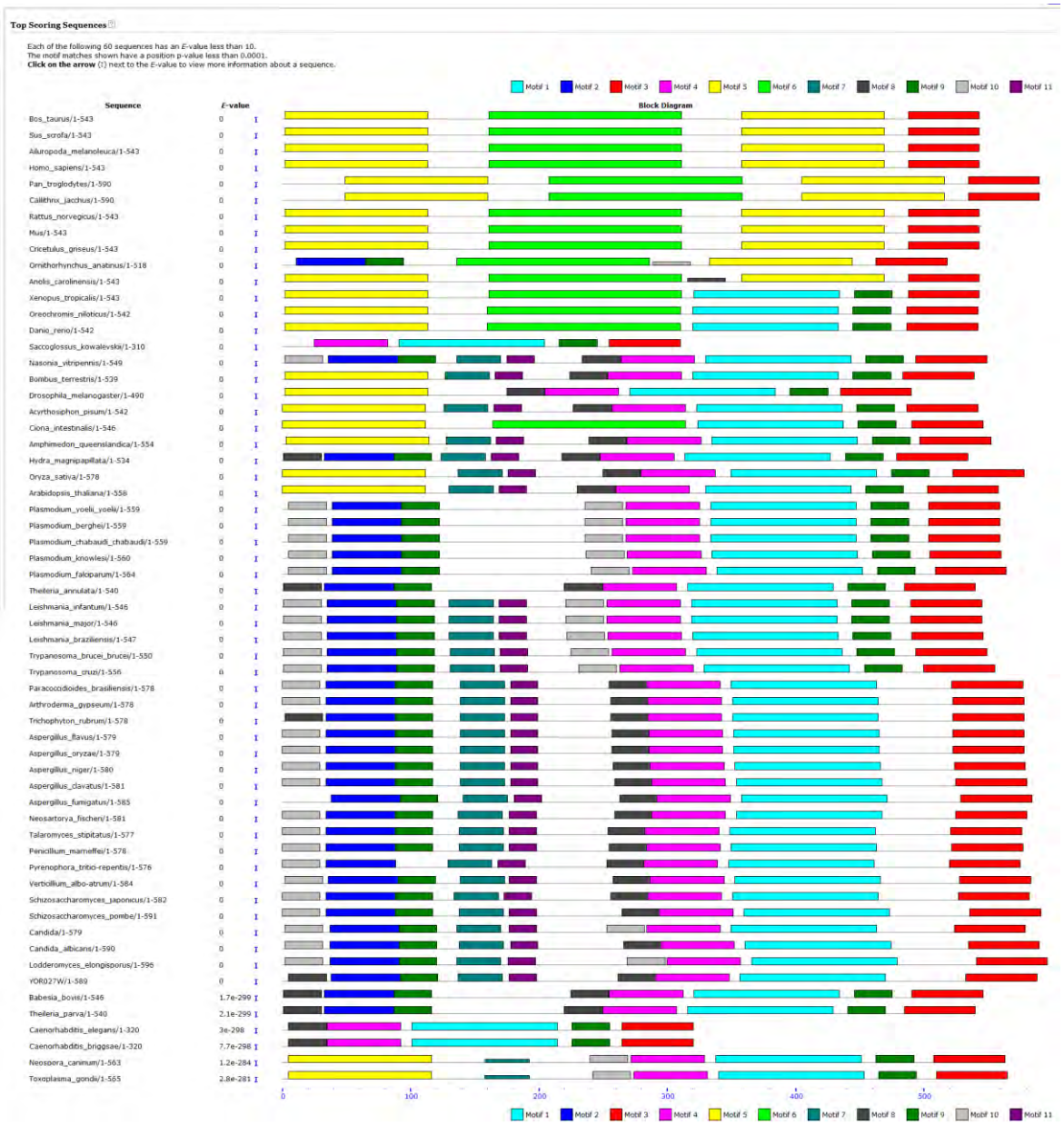


Figure A5.24: Mast results for all sequences submitted to Meme for 20 motif search

Appendix 6: Pairwise Alignments for HsHop, PfHop and ScHop

```
Score = 8090
Length of alignment = 593
Sequence S_cervisiae_YOR027W :  1 - 589 (Sequence length = 589)
Sequence        P_falciparum :  1 - 564 (Sequence length = 564)

S_cervisiae_YOR027W MSLTADEYKQQGNAAFTAKDYDKAIELFTKAIEVSETPNHVLYSNRSACYTS
                       |.  |. ||  |    |. |.  |. ||  ..  .|||||| |. ..|
       P_falciparum NKEEAQRLKELGNKCFQEGKYEEAVKYFSDAI-TNDPLDHVLYSNLSGAFAS

S_cervisiae_YOR027W LKKFSDALNDANECVKINPSWSKGYNRLGAAHLGLGDLDEAESNYKKALELD
                    | .| .||. || |. |.  |.||| | | |. || .| .||  |  .| .|
       P_falciparum LGRFYEALESANKCISIKKDWPKGYIRKGCAEHGLRQLSNAEKTYLEGLKID

S_cervisiae_YOR027W ASNKAAKEGLDQVHRTQQARQAQPDLGLTQLFADPNLIENLKKNPKTSEMMK
                    ..||. ...| .| | ..  .         ||.|. |   |.  |
       P_falciparum PNNKSLQDALSKV-RNENMLENA------QLIAHLN---NIIEN--------

S_cervisiae_YOR027W DPQLVAKLIGYKQNPQAIGQDLFTDPRLMTIMATLMGVDLNMDDINQSNSMP
                    ||||  |   ||..     ..|. .  . .| .  | . . .   .   |
       P_falciparum DPQL--KS--YKEENSNYPHELL-N-TIKSINSNPMNIRIILSTCHPKISEG

S_cervisiae_YOR027W KEPETS-KSTEQKKDAEP-QSDSTTSKENSSKAPQKEE-SKESEPMEVDEDD
                     |   . | | . .|||  | .     |   |  . |   || | |  ..
       P_falciparum VEKFFGFKFTGEGNDAEERQRQQREEEERRKKKEEEERKKKEEEEMKKQNRT

S_cervisiae_YOR027W -SKIEADKEKAEGNKFYKARQFDEAIEHYNKAWELH-KDITYLNNRAAAEYE
                     .|..| .|  || ||| ..||||. .|. | ... .|| |  |.|| . |
       P_falciparum PEQIQGDEHKLKGNEFYKQKKFDEALKEYEEAIQINPNDIMYHYNKAAVHIE

S_cervisiae_YOR027W KGEYETAISTLNDAVEQGREMRADYKVISKSFARIGNAYHKLGDLKKTIEYY
                    .|. |. |   |.|.  . .|..  ..| . |.. .| ..     .|| |
       P_falciparum MKNYDKAVETCLYAIENRYNFKAEFIQVAKLYNRLAISYINMKKYDLAIEAY

S_cervisiae_YOR027W QKSLTEHRTADILTKLRNAEKELKKAEAEAYVNPEKAEEARLEGKEYFTKSD
                    .||| |.         |.. |.   | | |||..|.|||| .  |.||| ..|
       P_falciparum RKSLVEDNNRATRNALKELERRKEKEEKEAYIDPDKAEEHKNKGNEYFKNND

S_cervisiae_YOR027W WPNAVKAYTEMIKRAPEDARGYSNRAAALAKLMSFPEAIADCNKAIEKDPNF
                    |||  |  |  |.|  |.||. ||||||||.||. .|  |.  |   |||| ||  |
       P_falciparum FPNAKKEYDEAIRRNPNDAKLYSNRAAALTKLIEYPSALEDVMKAIELDPTF

S_cervisiae_YOR027W VRAYIRKATAQIAVKEYASALETLDAARTKDAEVNNGSSAREIDQLYYKASQ
                    |.|| ||.  .. .|.|  ||.     |  |. |. . .. .|  . | . .
       P_falciparum VKAYSRKGNLHFFMKDYYKALQ----AYNKGLEL-DPNN-KECLEGYQRCAF

S_cervisiae_YOR027W QRFQPGTSNETPEETYQRAMKDPEVAAIMQDPVMQSILQQAQQNPAALQEHM
                    .  . . |.   || ....| |||.  |. ||  | |||. ..|| .. | .
       P_falciparum KIDEMSKSEKVDEEQFKKSMADPEIQQIISDPQFQIILQKLNENPNSISEYI

S_cervisiae_YOR027W KNPEVFKKIQTLIAAGIIRTG
                    |.| .|. .| ||||||..
       P_falciparum KDPKIFNGLQKLIAAGILKVR

Percentage ID = 34.23
```

Figure A6.1: Pairwise alignment of ScHop with PfHop, produced in Jalview.

Figure A6.2: Pairwise alignment of ScHop with HsHop, produced in Jalview.

Figure A6.3: Pairwise alignment of HsHop with PfHop, produced in Jalview.