Chapter 2

Structural Modelling of *P. falciparum* PPPK-DHPS

2.1. Introduction

To study the function of proteins it is vital to establish the coordinates of the important residues (preferably to atomic resolution) to gain insight into the mechanism of catalysis, substrate orientation and substrate binding. There are various methods available to obtain protein structure coordinates such as X-ray crystallography, cryo-electron microscopy and NMR measurements. These methods give various levels of resolution, with X-ray crystallography usually giving the highest resolution (down to 0.54Å (Jelsch *et al.*, 2000)) and cryo-EM the lowest (only down to secondary structure discrimination, 8-10Å). One of the Achilles heels of X-ray crystallography is the need to overexpress the protein or isolate sufficient quantities from natural sources. The latter method is a long and tedious process and thus overexpression in heterologous systems is the preferred method. Most proteins, especially malarial proteins, do not overexpress well in *E. coli* due to the rich A+T-codon bias (Aravind *et al.*, 2003). However, a protein may express in high enough yields but may not fold correctly and thus form inclusion bodies. Some proteins such as *P. falciparum* S-adenosyl-methionine decarboxylase, are expressed at low levels and are only stable in solution for a few hours, thus the risk increases for low, inactive protein yields (Birkholtz *et al.*, 2004). Once sufficient quantities of pure protein have been obtained it has to subjected to the procedure of crystallization which is regarded by most scientists as a "black box". Some effort has gone into trying to elucidate crystallization parameters, such as pI of the protein (Kantardjieff *et al.*, 2004) and the pH at which crystallization occurred (Dougall *et al.*, 2004), but they have met with limited success. In spite of these hurdles on the way to obtaining protein crystals, a resolution of 0.54Å is currently obtainable (Jelsch *et al.*, 2000).

NMR is one alternative option but suffers from some of the same problems as X-ray crystallography (expression levels, protein isolation). An advantage of NMR is that dynamic structures are obtained but currently only proteins up to ~80 kD can be investigated (Tugarinov *et al.*, 2002). In the early 70's researchers started to investigate computer models as an alternative to study protein structures.

The advantage of using computer modelling is the speed with which a protein can be modelled but the resolution achieved is dependent on the percentage identity between the target and template. The field of computer-aided protein modelling expanded during the 90's with the release of the MODELLER package (Sali *et al.*, 1993).

2.1.1. Homology modelling

Homology (or comparative) modelling is the process where the structure of a protein is modelled based on a closely related protein with a known structure. It is usually assumed that the proteins are homologous, meaning they have a common evolutionary ancestor. The 3D structures of proteins change little with evolution and thus the assumption which underpins comparative modelling is that proteins with similar sequences will have a similar structure (Marti-Renom *et al.,* 2000). The opposite is also true in that proteins with a similar structure do not necessarily share the same sequence (Marti-Renom *et al.,* 2000).

There are three main methods used to construct models: modelling by rigid-body assembly, modelling by segment matching and modelling by satisfaction of spatial restraints. Rigid-body assembly is the method whereby a protein model is constructed from small rigid pieces. These pieces are obtained from sequence alignments and are based on conserved core regions, loops and side-chain conformations. The backbone is constructed by taking the average C_α atom coordinates from the conserved part of the template structures. The rest of the main chain atoms are then superimposed from the template sequence with the highest similarity to the target. The loops are built after a library search for similar structures and the side-chains are built using a combination of template side-chain conformation and preferred side-chain conformation. The COMPOSER package (Sutcliffe *et al.*, 1987) applies this method of modelling.

Modelling by segment matching uses short peptide sequences to construct a model. By searching a database with the specific short sequence and matching C_α atom coordinates, short pieces of protein can be assembled. These short pieces are then put together and a conformational restraint is applied. The advantage of this method is that short insertions and deletions as well as side-chain atoms can be constructed. The method is implemented in the SEGMOD package (Levitt, 1992).

Spatial restraint satisfaction modelling is one of the most promising methods as various restraints can be used to guide protein construction. A program like MODELLER assumes that aligned residues will have similar restraints such as bond angles, dihedral angles and stereochemistry and uses these values to construct the model (Figure 2.1; Sali *et al.*, 1993).

The spatial restraints are then combined with the CHARMM22 forcefield parameters (to enforce proper stereochemistry) into an objective function (MacKerrell *et al.*, 1998). The model then undergoes optimization of the objective function in Cartesian space. The inclusion of the CHARMM22 forcefield

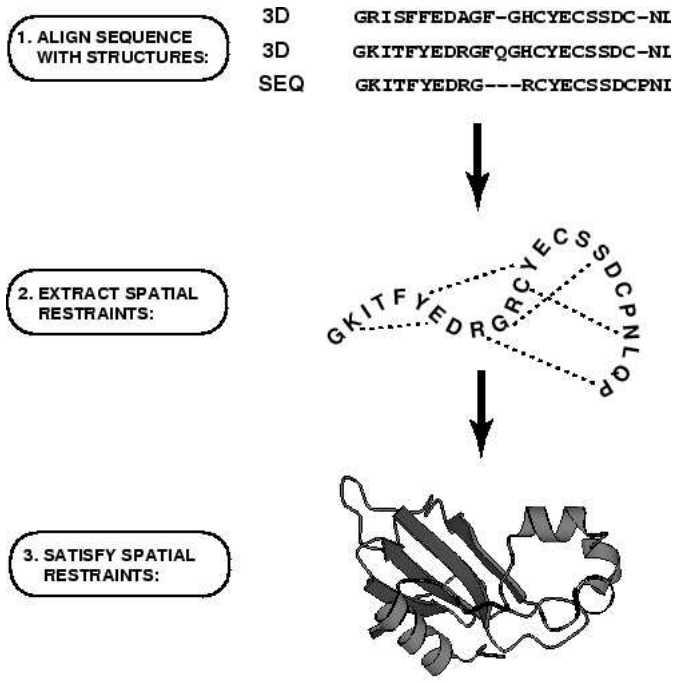


Figure 2.1: The basic methodology followed by MODELLER for comparative modelling by satisfaction of spatial restraints. The target sequence is aligned with the templates, the restraints extracted from the crystal structures and the model is built and optimised (from the MODELLER 6 manual, http://salilab.org/modeller/manual).

parameters allows for the extra optimization of the model through molecular dynamics. MODELLER also allows for the addition of extra restraints such as enforcing secondary structure, incorporating results from cross-linking experiments and NMR results, which can help to improve the model. Further discussion will relate to MODELLER, as it was used to construct the models in this study.

Loop modelling is one of the most difficult areas of homology modelling. Some enzymes differ only by the length of the loops between the secondary structure elements and these loops are more often than not also responsible for the specificity of the enzyme (Marti-Renom *et al.,* 2000). Various methods have been used to try and model loops such as database searches of known loop structures (Chothia *et al.*, 1987; Greer, 1980), molecular dynamics simulations (Bruccoleri *et al.*, 1990; Fiser *et al.*, 2000), genetic algorithms (McGarrah *et al.*, 1993; Ring *et al.*, 1994), biased probability Monte Carlo search (Evans *et al.*, 1995; Thanki *et al.*, 1997) and systematic conformational searches (Brower *et al.*, 1993; Bruccoleri *et al.*, 1993) to mention but a few. MODELLER once again tries to optimize the objective score of the loop to improve the model.

Another issue encountered when modelling sequences with low similarity is that of side-chain conformations. When two sequences have a 30% or less similarity, side-chain packing and interactions generally become unpredictable. Exposed side-chains also present a problem as they may adopt various different conformations without a single one being dominant (Chothia *et al.*, 1987; Rost, 1999).

As with any theoretical method there are certain restrictions during comparative modelling. Errors can be introduced in the protein backbone as a result of insertions (areas in the target without a template match) or deletions in the target sequence. Misalignment is one of the only errors which can be corrected to a certain extent. Sequence diversity will also introduce side-chain packing and interaction errors as well as using wrong templates in cases of distantly related proteins (Marti-Renom *et al.,* 2000).

Homology modelling (or comparative modelling) with MODELLER is divided into four steps: (1) fold assignment and template selection, (2) template-target alignment, (3) model building and (4) model evaluation and refinement. Template selection and fold assignment are some of the most important steps in comparative modelling. Various methods such as BLAST (Altschul *et al.,* 1990) and PSI-BLAST (Altschul *et al.*, 1997) can be used to identify templates from sequence and protein databases such as SCOP (Hubbard *et al.*, 1999) and the Protein Data Bank (PDB). When template-target similarity is low, methods such as threading or 3D template matching methods can be used to identify the fold. Threading methods rely on pairwise comparisons of the template and a protein structure with subsequent scoring of the resulting fit. When fold assignment is difficult due to low sequence similarity, a preliminary model can be built with threading methods. This preliminary model can then be evaluated based on known properties of the protein such as interaction sites, active site residues and conserved residues (Marti-Renom *et al.,* 2000).

To achieve a reliable model an optimal alignment between the templates and target is needed. Alignments are produced from the database hits when searching for templates but these are more often than not sub-optimal. To achieve optimal alignments and hence accurate models, the sequence alignment needs to contain as many related sequences as possible. This will ensure that the globally conserved regions are identified and that this information is incorporated into the template-target alignment. Modern methods such as Hidden Markov Models (HMM) have been used to create models of protein families to help in identifying more remote homologs (Grundy *et al.,* 1997). HMM's have been used to construct the Pfam database, which contains HMM's of all the known protein families (Bateman *et al.*, 2004). To create an HMM, a sequence alignment of all the known sequences are made and the resulting HMM is used to search for new, related proteins. By aligning the target with the alignment used in creating the specfic HMM, the globally conserved regions in the target can be identified. This helps in aligning the template and target in local regions where there is no or low sequence similarity. More modern methods such as 3D T-coffee use information contained in the crystal structure to help align the sequences (Poirot *et al.*, 2004). This ensures that in local, low similarity areas, alignment gaps are not introduced in structurally conserved regions. Another method which can help to improve alignments is secondary structure prediction. Running multiple secondary structure prediction algorithms on the template sequence will help in achieving a consensus of the predicted secondary structure. The predicted secondary structure can help to improve alignments in areas where there is low sequence similarity between target and template.

One or more templates can be used to build a model but in order to avoid small structural errors in the model only closely related structures should be used (if possible).

After the model has been built it needs to be evaluated. The first and most basic step is to check if the model has the correct fold. The wrong fold can usually be avoided by using a closely related template to start with. Once the fold has been confirmed, local areas can be analysed. If the sequence similarity is above 30% the correct folding of local areas and even some chains can be assumed (due to the higher sequence similarity). Once again low similarity (<30%) sequences will have to be checked for correct side-chain orientations. There are various programs which check some basic stereochemistry (PROCHECK, Laskowski *et al.*, 1993; WHATCHECK, Hooft *et al.*, 1996). Lazaridis and Karplus (Lazaridis *et al.*, 1999) used the CHARMM vacuum potential together with a Gaussian distribution model for solvation free energy to show that correctly folded states are always more stable than misfolded states. Their conclusion was that a molecular mechanics energy function coupled to a simple solvation free energy model can perform as well as statistical methods to discriminate between correctly and incorrectly folded proteins. For this study we used PROCHECK.

There are a few limitations to homology models. The resolution of the models cannot go down to atom level (except possibly in high homology cases where sequence similarity is >70%). Another limitation is that homology modelling cannot yet do *ab initio* modelling of sequences longer than 7 residues. The next section will describe the implementation of the discussed methods on the *P. falciparum* PPPK and DHPS sequences.

2.2. Methods

Several steps need to be followed to ensure an accurate homology model of a protein. The first is a correct multiple sequence alignment for all the known protein sequences from the desired family. This ensures that any globally conserved regions are identified. The multiple sequence alignment is then used to achieve an accurate template-target match for the modelling process. The second step is to predict the secondary structures of any regions which do not have a template match. Incorporating the predicted secondary structure can help in increasing the overall usefulness of the model. A third step is to include co-crystallized ligands and co-factors in the model. This helps to restrain the active site residues to the correct conformations. The application of these steps to *P. falciparum* PPPK and DHPS will be discussed in the following sections in more detail.

2.2.1. Homology modelling of *P. falciparum* DHPS

2.2.1.1. Plasmodial DHPS Sequence Retrieval

A Swiss-Prot/TrEMBL (http://www.expasy.org/sprot/, Boeckmann *et al.*, 2003) search for Plasmodial DHPS revealed partial sequences for two members of the *Plasmodium* family (*P. falciparum* and *P. chabaudi*). Both of these entries have been classified as partial sequences (they have been translated from genomic sequences and have not been confirmed) but they seem to contain the complete DHPS sequence. The *P. falciparum* DHPS sequence was used as the target sequence for modelling (TrEMBL accession nr: Q27865) and the *P. chabaudi* sequence (TrEMBL accession nr: Q9BLN5) was added to aid in the multiple sequence alignment. To obtain sequences not yet in Swiss-Prot, a tblastx and blastp search was done against the NCBI genome database (http://www.ncbi.nlm.nih.gov/Blast.cgi). Tblastx is a modified form of BLAST, which takes a DNA query and translates it into all six reading frames and then queries the results against a translated database. From the tblastx results of the DNA sequence of *P. falciparum* DHPS, the sequence of *P. berghei* DHPS (Genbank accession nr: AF344661) was found. From the blastp search the sequence of *P. yoelli yoelli* (Genbank accession nr: EAA21661) was also obtained. Towards the end of the project (2004) Korsinczky *et al.* (2004) published the sequence of the *P. vivax* PPPK-DHPS gene (Genbank accession nr: AY186730). This sequence was not included in the orignal alignment, however, it was used later during the studies of the parasite-specific insertions and their conservation.

2.2.1.2. Multiple Sequence Alignment

An accurate multiple sequence alignment of DHPS sequences is needed in order to deduce the correct conserved regions between all species. Pfam (Bateman *et al.*, 2004) is a large database of protein family alignments and Hidden Markov Models (HMM) derived from the multiple sequence alignments of all the known sequences for protein families. Pfam provides the HMM as well as a choice between the complete multiple sequence alignment of the family or a core multiple sequence alignment (also called a seed alignment) of the family. By using the complete multiple sequence alignment of the DHPS family as a profile, the Plasmodial sequences can be aligned according to conserved positions in all the known sequences. This usually results in very accurate multiple sequence alignments, which incorporate structural conservation (due to the large number of sequences used for the complete multiple sequence alignment).

The *M. tuberculosis* DHPS (MtbDHPS) crystal structure was used as one of the templates for homology modelling as it had the highest identity (30%) to *P. falciparum* DHPS, as well as a vital Mg^{2+} ion in

the active site. *B. anthracis* DHPS (BaDHPS) was also used as it contained a product analogue in the active site. The product analogue crystallized into *B.anthracis* DHPS provided the first concrete clues to the location of *p*-aminobenzoic acid in the DHPS active site. The profile alignment mode in ClustalX (Thompson *et al.*, 1994) was used to align the Plasmodial sequences to the DHPS seed alignment from Pfam (PF00809). The resulting alignment was used to align the templates (*M. tuberculosis*, *B. anthracis*) and the target sequence (*P. falciparum*) for homology modelling according to globally conserved features. To help ensure the accuracy of the alignment the sequences were also submitted to MEME (Bailey *et al.*, 1994) to identify conserved motifs throughout all the known DHPS sequences.

2.2.1.3. Secondary Structure Prediction

Proteins usually consist of a combination of α-helices and β-sheets. The secondary structures of the globally conserved regions in *P. falciparum* DHPS were already known based on the alignment but the inserts had no known template match. The *P. falciparum* DHPS sequence was submitted to various secondary structure prediction servers. These included the DSC secondary structure prediction server (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_dsc.html, King *et al.*, 1996), the PredictProtein server (http://cubic.bioc.columbia.edu/predictprotein/, Rost *et al.*, 2003), PSIPRED server (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html, McGuffin *et al.*, 2000), 3-D Jury server (http:// bioinfo.pl/Meta/, Ginalski *et al.*, 2003) and the HMMSTR/Rosetta server (http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php, Bystroff *et al.*, 2000). All the results from the various servers were combined to obtain a consensus result for the secondary structure of the inserts.

To obtain a better understanding of the parasite-specific inserts as well as the overall structure of PfDHPS, the program SEG (Wootton *et al.*, 1993) was used for low complexity prediction and ProMate (Neuvirth *et al.*, 2004) was used for protein-protein interaction predictions. In this study the inserts were defined as those areas which did not show any alignment matches in the global Pfam alignment.

2.2.1.4. Homology modelling

MODELLER 6v3 (http://salilab.org, Sali *et al.*, 1993) was used to build the homology models. For a first run the DHPS crystal structures of *M. tuberculosis* (Baca *et al.*, 2003, PDB code: 1EYE), *S. aureus* (Hampele *et al.*, 1997, PDB code: 1AD1) and *E. coli* (Achari *et al.*, 1997, PDB code: 1AJ2) were used as templates (the *B. anthracis* structure was not yet available at this time). The result of the multiple template usage was not optimal as some of the crystal structures contained ligands in the active site and some did not. Not all the crystal structures contained all the loops and thus it was later decided to use both *M. tuberculosis* and *B. anthracis* (Babaoglu *et al.*, 2004, PDB code: 1TX0) as templates. This ensured the least amount of residues without a template match while combining a product analogue and a

metal ion in the active site. The templates and the target were aligned according to the multiple sequence aligment obtained in section 2.2.1.2. Ten models (with and without the longer insertion, to investigate differences in regions lacking templates) were then built with MODELLER (hetero atom set to 'ON', water set to 'OFF'), which included the substrate analogue and Mg^{2+} ion from the crystal structure. This ensured that the restraints imposed on the active site by the crystal structure were transferred to the model during model building. In an effort to improve accuracy the extended loop modelling function of MODELLER was used. The validity of the models was judged by using the MODELLER objective function as well as the overall structure of the model. The model with the longer insert only resulted in a peptide chain extending from the model and thus the model without the longer insertion was used for further study. The model with the highest objective score, which did not contain any knots in the protein backbone (knots are where the backbone of the protein forms a loop and then the backbone pass through the loop), was chosen. The objective score is a function implemented by MODELLER to assess models. The objective score function is based on molecular dynamics principles and the program tries to optimise certain parameters for the model. MODELLER was run on a dual Pentium III (1 GHz) cpu, 1GB RAM, running SuSE 9.1.

2.2.1.5. Mutation Modelling

There are at least five known mutations linked with resistance of sulfadoxine to *P. falciparum* DHPS (S436A/F, A437G, K540E, A581G, A613S/T, Triglia *et al.*, 1997). In order to investigate the effect of the mutations, InsightII (Accelrys) was used to produce the mutation-containing models. The original model was loaded into InsightII and then, using the Biopolymer module, was modified to the desired mutations. The following mutation models were constructed:

- S436A/F
- A437G
- K540E
- A581G
- A613T

These models were subjected to 100 steps of minimization (to ensure that no major energy clashes occurred) in InsightII and further used to investigate the effect of the different mutations on substrate and drug orientation in the active site. InsightII was run on a Silicon Graphics Inc. Octane2 workstation with 512MB RAM.

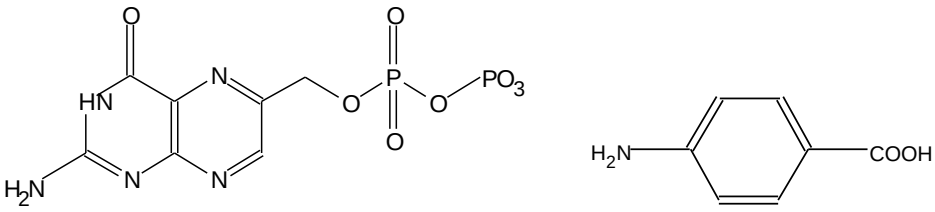


Figure 2.2: The structure of 6-hydroxymethyl-7,8-dihydropterin pyrophosphate (DHP) (left) and *p*-aminobenzoic acid (right). The amino group on *p*-aminobenzoic acid attacks the carbon atom linking the pyrophosphate group and the ring system on DHP (Baca *et al.*, 2000).

2.2.2. Ligand Docking in *P. falciparum* DHPS

The *M. tuberculosis* crystal structure contained a substrate analogue and Mg^{2+} ion, and *B. anthracis* contained a product analogue. To obtain a global picture of ligand-protein interaction, the orientation of the natural substrates and sulfadoxine in the wild-type and the resistant types needed to be compared. This was achieved by building the ligands into the model from the crystal structures with MODELLER, or docking them into the active site model using docking software such as the Ligand Fit module in Cerius2 (Accelrys). The product analogue was modified to the two natural subtrates (using InsightII) and sulfadoxine was docked into the active site using Cerius2 (Accelrys).

2.2.2.1. Substrate Construction

During the modelling process, the Mg^{2+} ion in *M. tuberculosis* DHPS and the product analogue in *B. anthracis* DHPS were incorporated into the model for PfDHPS. The model was loaded into InsightII and the product analogue was modified to the natural substrates, 6-hydroxymethyl-7, 8-dihydropterin pyrophosphate (DHP) and *p*-aminobenzoic acid (Figure 2.2). The substrates were saved as .pdb files for later use in ligand docking.

2.2.2.2. Sulfa-drug Construction and Docking

Various sulfa-drugs are in use against parasites but the most commonly used drug against malaria is sulfadoxine (Figure 2.3). Most of the sulfa-drugs are *p*-aminobenzoic acid analogues with various functional groups attached. The following sulfa-drugs were constructed with the Builder module of InsightII and saved as .mol2 files for later use: sulfathiazole, dapsone, sulfachloropyridazine, sulfaquinoxaline, sulfadimethoxine, sulfapyridazine, sulfamoxole, sulfamethoxypyridazine, sulfamethoxazole, sulfamerazine, sulfadiazine, sulfisoxazole, PAS, sulfanilamide and sulfacetamide (Kasekarn *et al.*, 2004). The constructed sulfa-drugs were minimized in InsightII until the change in energy was less than 0.01 kcal/mol using the CVFF forcefield (Accelrys).

Only sulfa-drug resistance with respect to sulfadoxine was investigated and not the detailed docking of the other sulfa-drugs. For the docking of the sulfa-drugs, the LigandFit module of Cerius2 (Accelrys) was

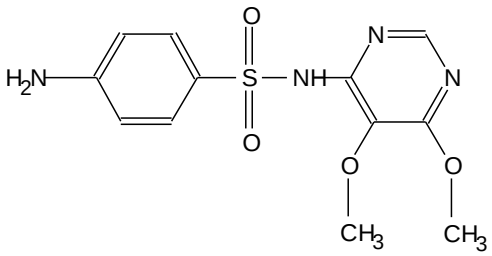


Figure 2.3: The structure of sulfadoxine, a *p*-aminobenzoic acid analogue (Kasekarn *et al.*, 2004).

used. The model was loaded with the substrates in place (as constructed in InsightII). The Cavity Finding module was used to identify all the cavities and the active site cavity was selected. The following programs were also used in attempts to dock sulfadoxine: the LUDI module of InsightII, Autodock (Morris *et al.*, 1998) as well as GOLD (Jones *et al.*, 1997). Cerius2 LigandFit produced 20 fits with a range of conformations. The orientation in which the *p*-aminobenzoic acid part of sulfadoxine showed the closest resemblance to the crystal struture orientation was chosen. Only Cerius2 succeeded in docking sulfadoxine in an acceptable conformation and orientation. An acceptable conformation was a conformation in which sulfadoxine would interact with DHP in a way similar to the PfDHPS catalytic mechanism.

The following parameters were used and applied during the docking of sulfadoxine into the DHPS model with Cerius2:

- Hydrogens were automatically checked to ensure that all residues had the correct number of hydrogens.
- The automatically identified binding pocket was enlarged to ensure an adequate sampling of the conformational space in the active site.
- The default grid-resolution size was kept at 0.5Å to give an accurate representation of the active site.
- A pore opening of 8Å was specified to ensure that an adequate region of conformational space was sampled.
- The option to penalise any orientation of sulfadoxine if some part of the molecule was beyond the defined active site, was activated.
- Cerius2 was allowed to make the ligand flexible.

The sulfadoxine-PfDHPS complex was minimized according to the same criteria used in section 2.2.2.2 for minimizing the ligand-PfDHPS complex. The interaction between the DHPS model and the substrates was then visualized using LIGPLOT (Wallace *et al.*, 1995). The interaction between DHPS and sulfadoxine was compared to the interactions published for *M. tuberculosis, B. anthracis, S. aureus* and

E. coli DHPS (Baca *et al.*, 2000). Cerius2 was run on a Silicon Graphics Inc. Octane2 workstation with 512MB RAM.

2.2.3. Homology modelling of *P. falciparum* PPPK

2.2.3.1. Multiple Sequence Alignment

To ensure a reliable alignment of PfPPPK with the template, the PPPK (Pfam reference: PF01288) profile from Pfam was used. This ensured a reliable alignment of the globally conserved regions. The complete *P. falciparum* PPPK sequence from Genbank (Genbank accession nr: U07706) was used as well as partial sequences from *P. vivax* (Genbank accession nr: AY186730), *P. berghei* (Genbank accession nr: AF344661) and *P. chabaudi* (Genbank accession nr: AJ302077), which were obtained through a tblastx search with the *P. falciparum* PPPK DNA sequence. Blastp provided the PPPK sequence of *P. yoelli yoelli* (Genbank accession nr: EAA21661). The globally conserved regions were used to align the sequences for which a crystal structure was available. PPPK of *E. coli* (PDB ref: 1DY3, 1CBK, 1EQM, 1EQO, 1EX8, 1F9H, 1F9Y, 1G4C, 1HKA, 1HQ2, 1IM6 and 1KBR) and *H. influenzae* (PDB ref: 1CBK) had available crystal structures. PfPPPK showed 21% identity and 40% similarity to EcPPPK.

2.2.3.2. Homology modelling

A reliable structural model needs to include as many restraints as possible when building the model. All the available crystal structures of PPPK were investigated. There are 12 PPPK crystal structures available (11 from *E. coli* (Ec; Stammers *et al.*, 1999; Blaszczyk *et al.*, 2000) and one from *H. influenza* (Hi; Henning *et al.*, 1999)). The *E. coli* crystal structure was used as it contained the most complete set of substrates (ATP and a pterin substrate analogue, 7,8-dihydro-6-hydroxymethyl-7-methyl-7-pterin) and co-factors (two Mg^{2+} ions). The other crystal structures (PDB ref: 1CBK, 1EQM, 1EQO, 1EX8, 1F9H, 1F9Y, 1G4C, 1HKA, 1HQ2, 1IM6 and 1KBR) contained various forms of substrate analogues and either ADP or AMP. From the multiple sequence alignment it was evident that *P. falciparum* PPPK contained two large inserts (each about 90 residues in length when compared to non-Plasmodial species) and the decision was made to exclude them from the model building process due to their length and the lack of a template match. MODELLER 6v3 was used to build the models. The substrate analogue (7,8-dihydro-6-hydroxymethyl-7-methyl-7-pterin) and ATP, as well as the two Mg^{2+} ions from the crystal structure, were included during the model building. The models were built following the same criteria used for the modelling of *P. falciparum* DHPS in section 2.2.1.4. The substrate analogue

was converted to the natural pterin substrate using InsightII. To obtain a better characterization of the protein, low-complexity region prediction was done using the program SEG (Wootton *et al.*, 1993).

2.3. Results

The following results were obtained from the homology modelling of PfDHPS and PfPPPK.

2.3.1. Homology Modelling of DHPS

The template-target alignment was derived from the Pfam DHPS multiple sequence alignment and the additional four Plasmodial species were added later. Figure 2.4 shows *P. falciparum* DHPS aligned with the four other Plasmodial DHPS sequences as well as the *M. tuberculosis* and *B. anthracis* DHPS sequences. Notice the two inserts when comparing the five Plasmodial species to *M. tuberculosis* and *B. anthracis*. The length of the inserts also differ between the five Plasmodial species but there is a high level of conservation between the Plasmodial species from residues 740-753 in the second insert. The length of insert 1 varies between the Plasmodial species and it does not seem to have any conserved regions. The alignment shows that there is a highly conserved, *Plasmodium*-specific region at the end of insert 2 (residues 747-759 in the *P. falciparum* sequence), which was predicted to be an α-helix. Figure 2.5 shows the conserved motifs in known DHPS sequences predicted by MEME. The alignment suggested that there are two inserts when comparing the DHPS of the Plasmodial species to the rest of the known DHPS sequences. The inserts in the *P. vivax* and *P. falciparum* sequence are shown in Figure 2.5. Motif 13 seems to be missing in *P. falciparum* but this is due to an insert in the junction region in the *P. falciparum* sequence.

Low complexity prediction using SEG, showed no low complexity regions in the *P. falciparum* DHPS sequence. Modeller produced 10 models and the best was chosen according to the criteria listed in section 2.2.1.4. The DHP and the Mg^{2+} ion were built directly into the model from the crystal structure. Figure 2.6 shows interactions in the DHPS model active site. The primary active site residues show a very high degree of conservation when compared to the available crystal structures (Table 2.1). An advantage of using homology modelling is that the active site residues of the model are restrained by the orientation of the template residues, thus providing a very accurate representation of the orientation of the active site residues. DHPS occurs as a TIM barrel and the model also assumed this fold. Figure 2.7 shows the comparison between the *P. falciparum* model, *M. tuberculosis* and *B. anthracis* crystal structures. The main difference between the two models can be seen in the loop regions where *P. falciparum* has loops which are a few residues longer. Figure 2.8 shows the orientation of the substrate, DHP, and the Mg^{2+} ion in the model. Note the proximity of the amino group of *p*-aminobenzoic acid to the carbon

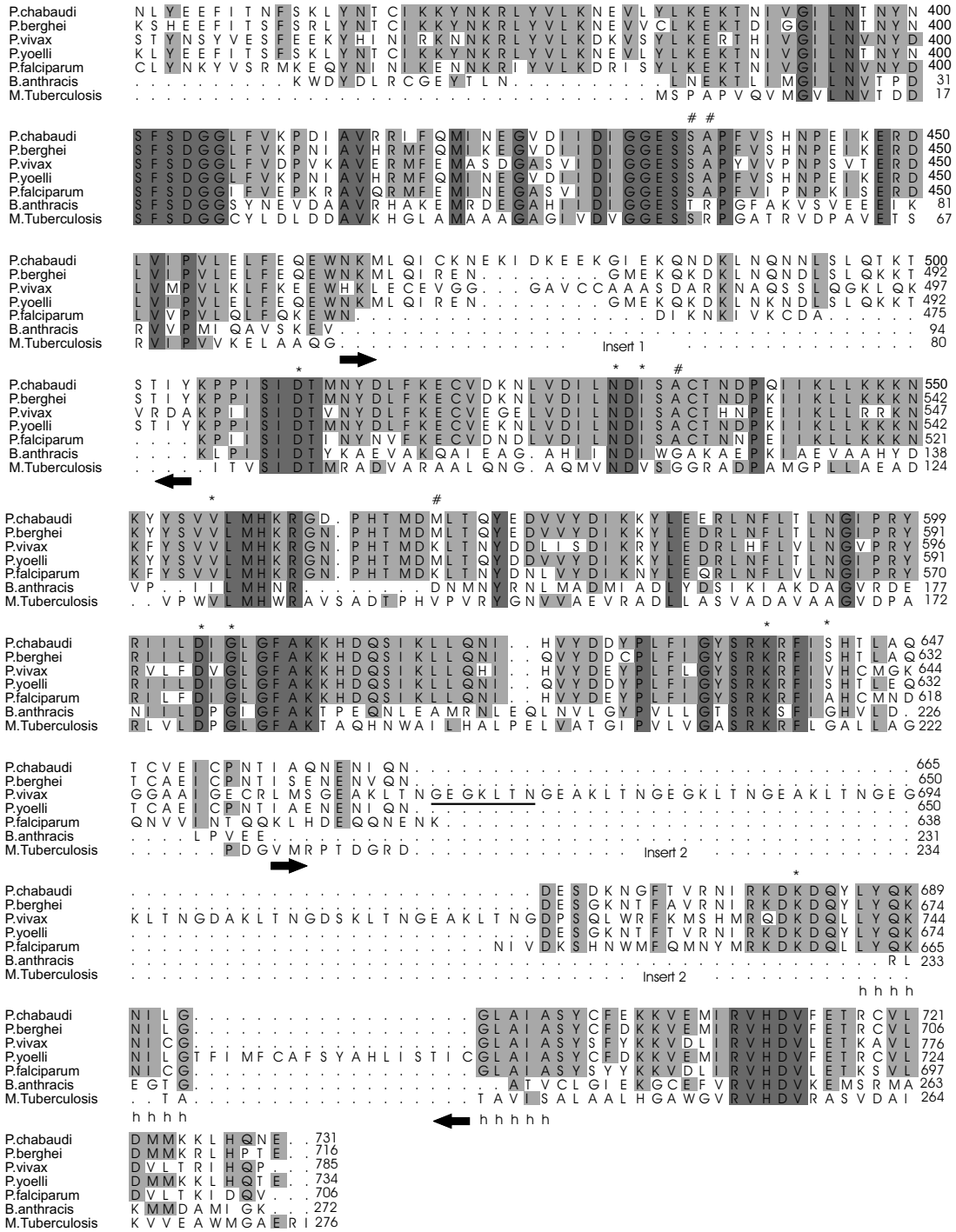


Figure 2.4: A multiple sequence alignment between *B. anthracis*, *M. tuberculosis* and five Plasmodial DHPS sequences (*P. chabaudi*, *P. berghei*, *P. yoelli yoelli*, *P. falciparum* and *P. vivax*). As DHPS forms part of the PPPK-DHPS gene, the DHPS sequence numbers for the *Plasmodium* species start at 350, thus maintaining the original numbering for the bifunctional protein. The arrows indicate the start and end of the insertions, the underlined region is the repeat present in *P .vivax*, 'h' indicates the predicted α-helix in insert 2, '#' indicates the known resistance-causing mutations and '*' indicates residues involved in direct substrate contact (Baca *et al.*, 2000).

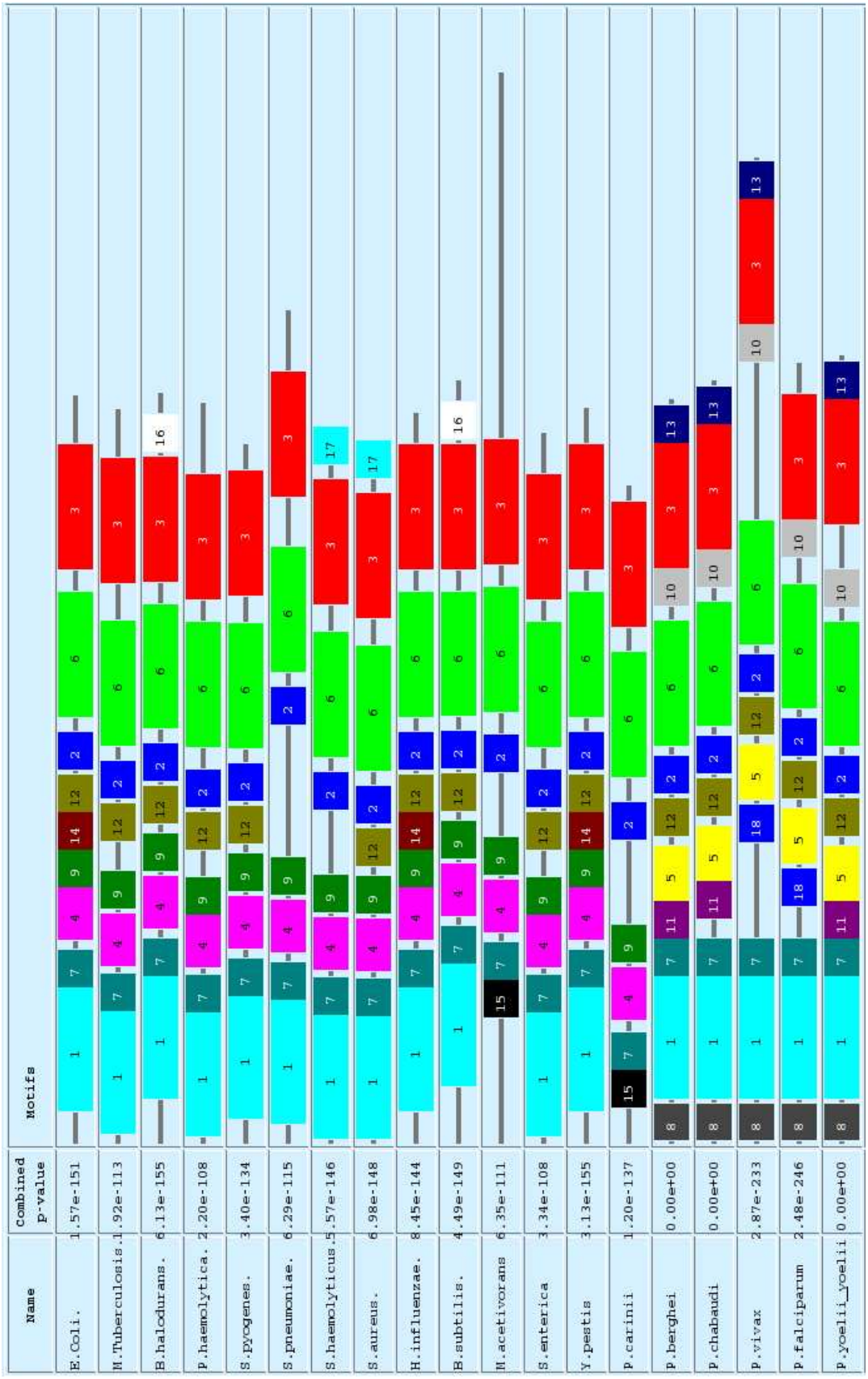


Figure 2.5: The conserved motifs in DHPS identified by MEME. Note the extra motifs 5, 8 and 10 identified by MEME. These motifs are unique to the *Plasmodium* species. The *Plasmodium* species also lack motif 4 and 9. Note the extra motif 11 in the sequences of *P. berghei*, *P. chabaudi* and *P. yoelli yoelli*. Motif 11 supports the hypothesis that these three species are closely related (Kedzierski *et al.*, 2002). *P. falciparum* also lacks motif 13 as it seems to contain an insert in the junction region when comparing it to the other *Plasmodium* species.

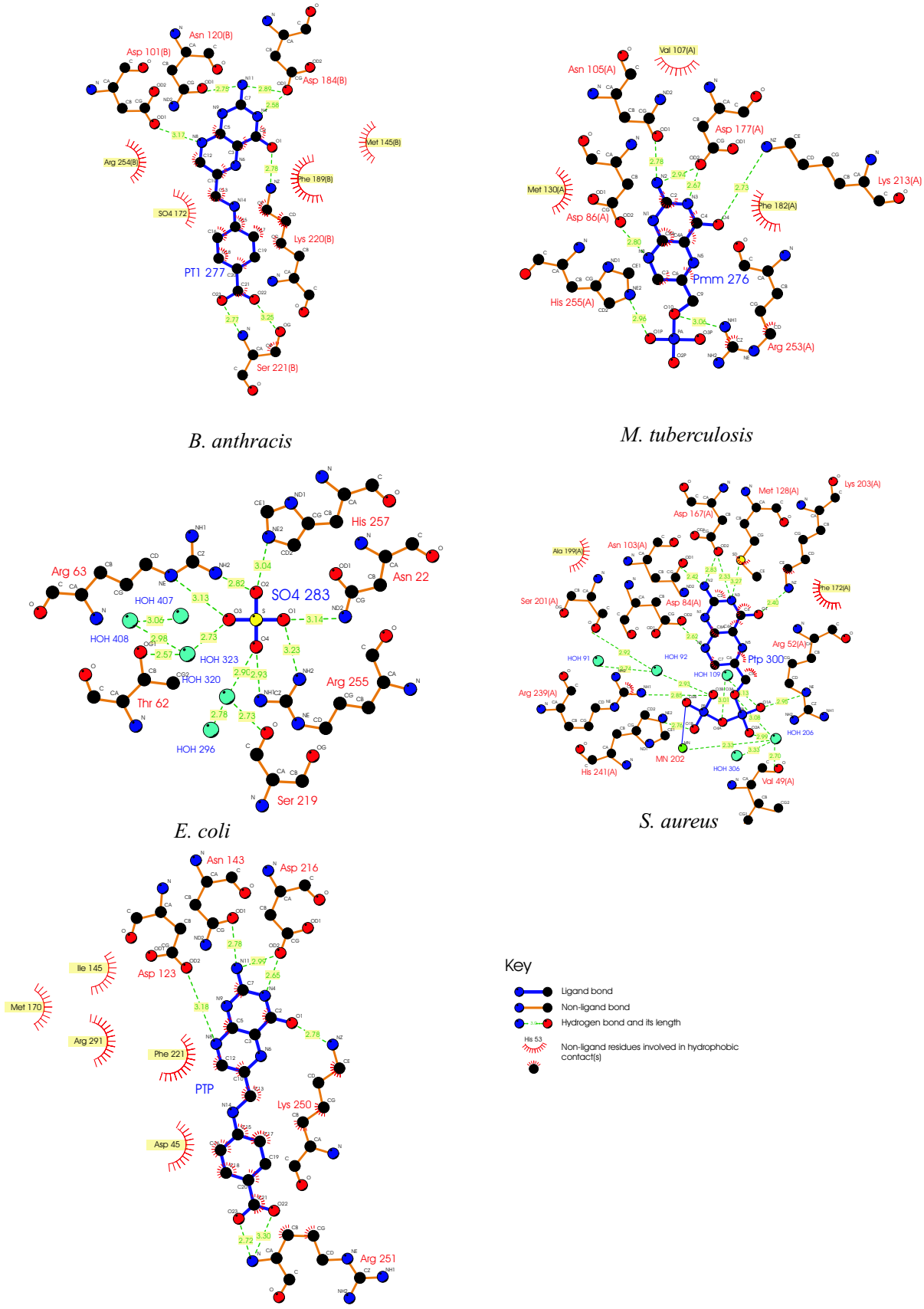


Figure 2.6: Contacts in the active site of the *P. falciparum* DHPS model in comparison to the known DHPS crystal structures. PT1, PTP and PMM indicate the positions of the substrate (or its analogues) in the known crystal structures. The numbering are those in the crystal structures (or in the case of *P. falciparum*, the model), PT1: pteroic acid, Pmm: pterin-6-yl-methyl-monophosphate, SO_4: sulfate ion, Ptp: pterin pyrophosphate. All figures produced by LIGPLOT (Wallace *et al.*, 1995).

connecting the pyrophosphate group to the ring system of DHP. During catalysis the amino group attacks this carbon and condenses with DHP. This causes the pyrophosphate group to be released from DHP, resulting in the formation of dihydropteroate. The model contained 9 α-helices and 8 β-sheets similar to the 9 α-helices and 8 β-sheets in the *M. tuberculosis* and *B. anthracis* crystal structures.

The PROCHECK quality check of the PfDHPS model is shown in Figure 2.9. The results for the Ramachandran plot and the estimated accessibility plots are shown. The model showed three residues in the disallowed regions. PROCHECK were also ran on both templates and compared to the model.

Table 2.1: Comparison of the active site residues between *M. tuberculosis*, *S. aureus*, *E. coli*, *B. anthracis* and the *P. falciparum* DHPS model. These residues are highly conserved in the alignment (Figure 2.4) (Baca *et al.*, 2000; Hampele *et al.*, 1997; Achari *et al.*, 1997; Babaoglu *et al.*, 2004).

M. tuberculosis	***S. aureus***	***E. coli***	***B. anthracis***	***P. falciparum****	**Contacts**
Asp 21	Asp 19	Asp 30	Asp 35	Asp 45/404	Interacts with the α-phosphate group on PTP
Asp 86	Asp 84	Asp 96	Asp 101	Asp 123/482	Has hydrogen bonds with the amino group on PTP
Asn 105	Asn 103	Asn 115	Asn 120	Asn 143/503	Has hydrogen bonds with the N-atoms in PTP
Met 130	Met 128	Met 129	Met 145	Met 170/529	Shows hydrophobic interactions with PTP
Asp 177	Asp 167	Asp 185	Asp 184	Asp 216/575	Hydrogen bonds with the amino group on PTP
Phe 182	Phe 172	Phe 188	Phe 189	Phe 221/580	Undergoes hydrophobic interaction with PTP
Ser 211	Ser 201	Ser 219	Ser 221	Ser 248/607	Interacts with *p*-aminobenzoic acid
Lys 213	Lys 203	Lys 221	Lys 220	Lys 250/609	Interacts with the α-phosphate group through a water molecule, as well providing hydrogen bonds to the 4-oxo group on PTP
Arg 253	Arg 239	Arg 255	Arg 254	Arg 291/686	Interacts with the α-phosphate group on PTP
His 255	His 241	His 257	His 256	His 293/688	Interacts with the α-phosphate group on PTP

* The first number is the residue number as it appears in the model. The second number indicates the residue number in the full bifunctional protein.

2.3.2. Ligand Docking

Sulfadoxine was docked into the active site of PfDHPS with the LigandFit module of Cerius2. The *B. anthracis* crystal structure revealed a probable position of *p*-aminobenzoic acid. As sulfadoxine is a *p*-aminobenzoic acid analogue, the *p*-aminobenzoic acid position helped in eliminating some of the possible orientations suggested by Cerius2. Figure 2.10 shows the orientation of sulfadoxine in the active

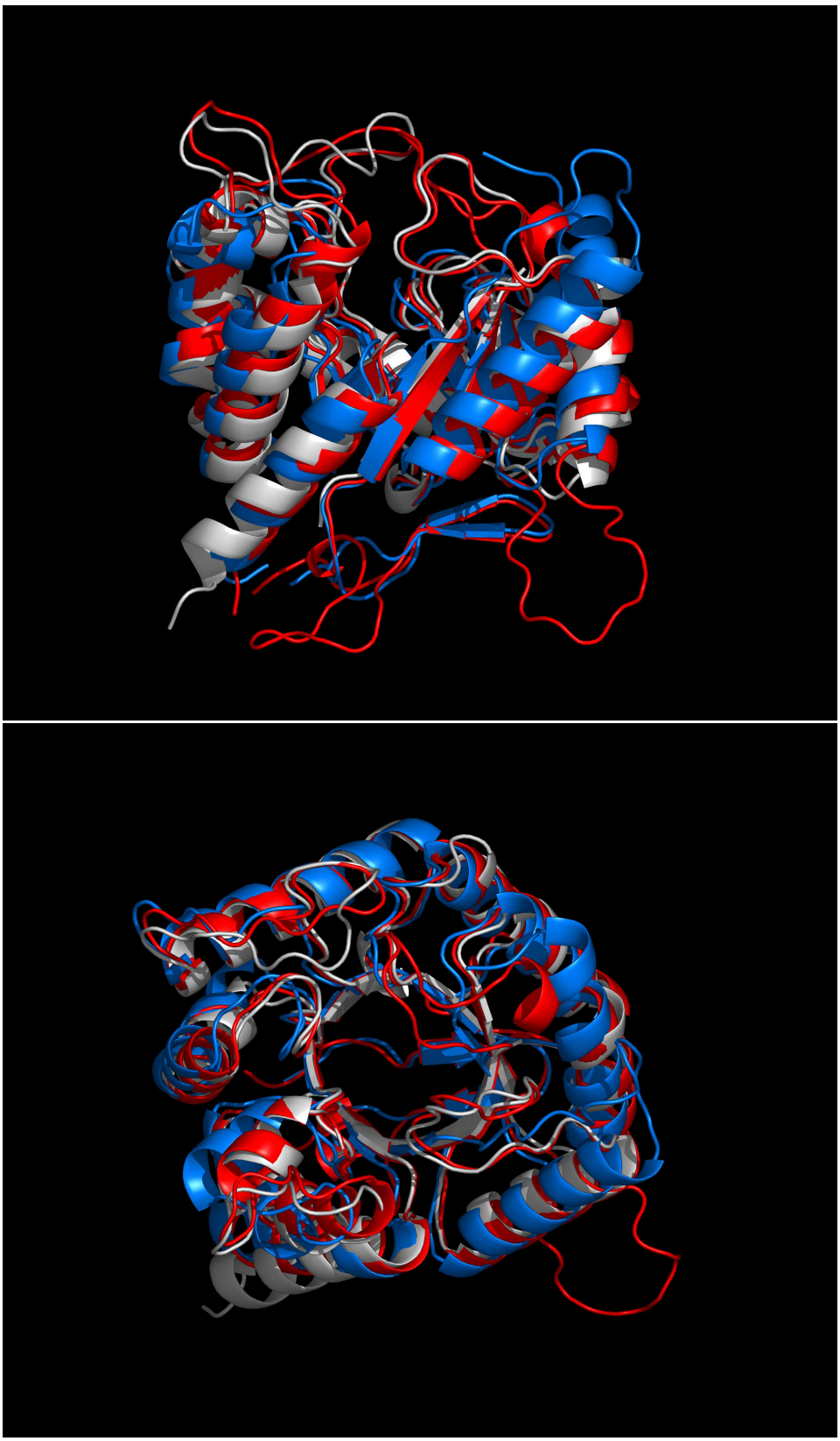


Figure 2.7: A comparison between the backbone structures of *M. tuberculosis* (white), *B. anthracis* (blue) and the *P. falciparum* (red) model. Top: side view of the DHPS model. Bottom: view from the top of the DHPS model. The red loop portruding from the bottom set of models is the shorter insert in the *P. falciparum* sequence (residues 463 to 475 in the bifunctional protein, residues 105 to 117 in the model). Generated with PYMOL (DeLano 2002, http://www.pymol.org).

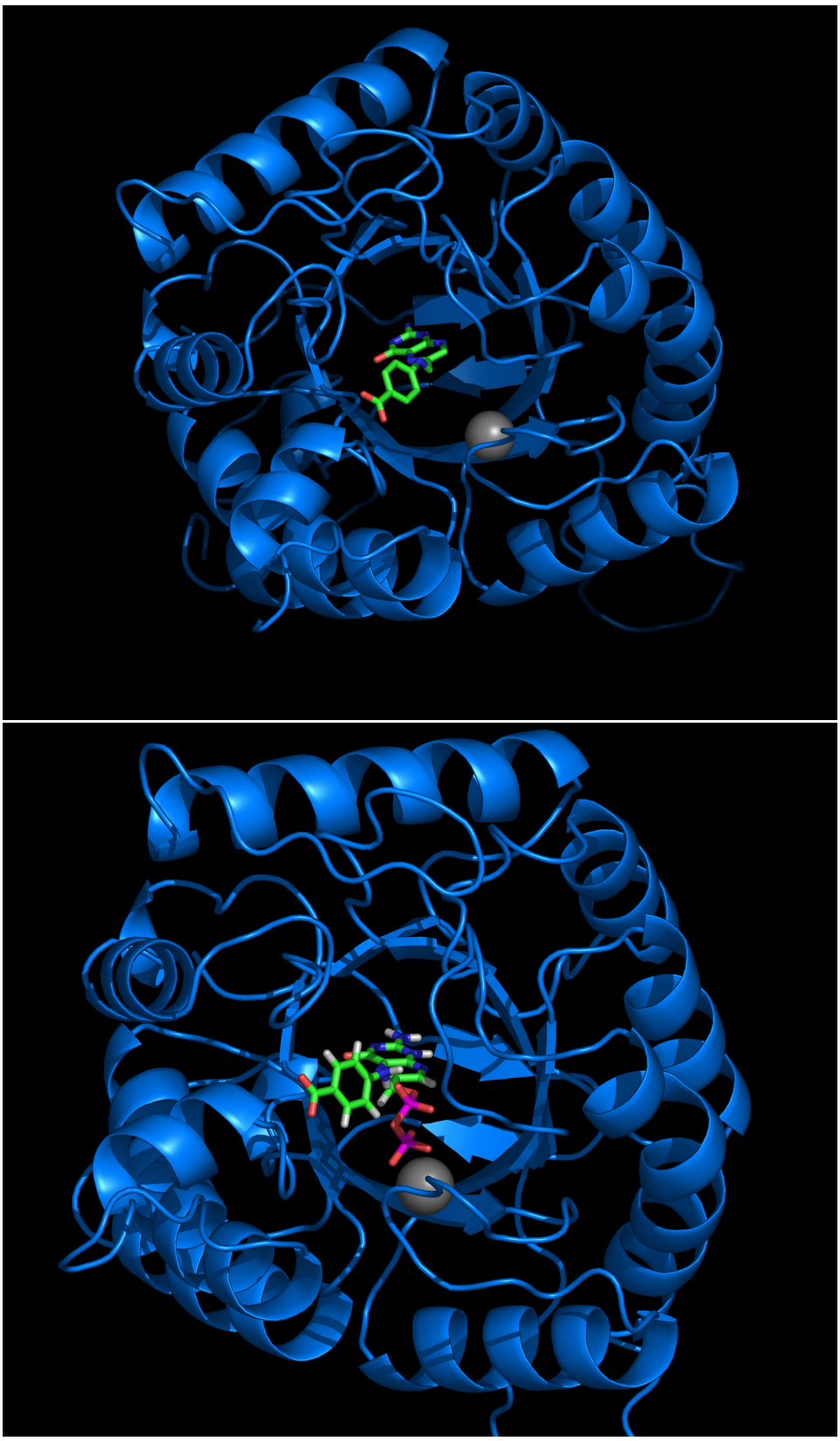
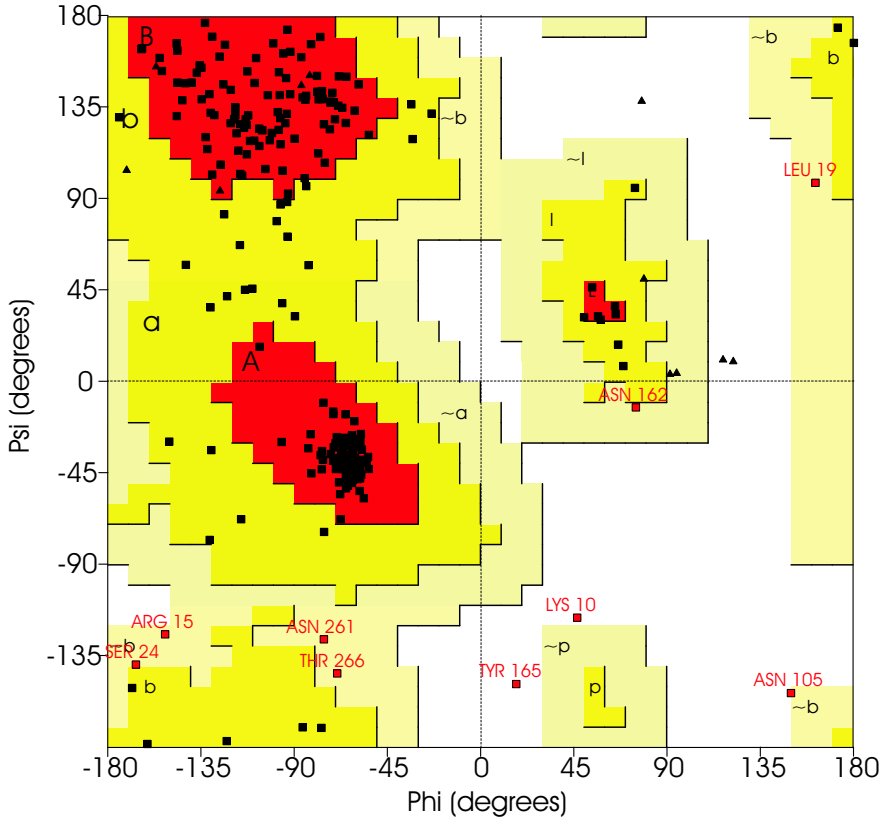


Figure 2.8: The *P. falciparum* DHPS model showing the orientation of the subtsrates and the Mg^{2+} ions. Top: The location of the product analogue and the Mg^{2+} ion (grey) after model construction. Bottom: The DHP substrate and Mg^{2+} ion were built in from the crystal structure during model building, and the *p*-aminobenzoic acid constructed by breaking the bond between the pterin group and the benzene moiety. Generated with PYMOL (DeLano 2002, http://www.pymol.org).



Plot statistics

Residues in most favoured regions [A,B,L]	240	83.9%
Residues in additional allowed regions [a,b,l,p]	37	12.9%
Residues in generously allowed regions [~a,~b,~l,~p]	6	2.1%
Residues in disallowed regions	3	1.0%
Number of non-glycine and non-proline residues	286	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	13	
Number of proline residues	10	
Total number of residues	311	

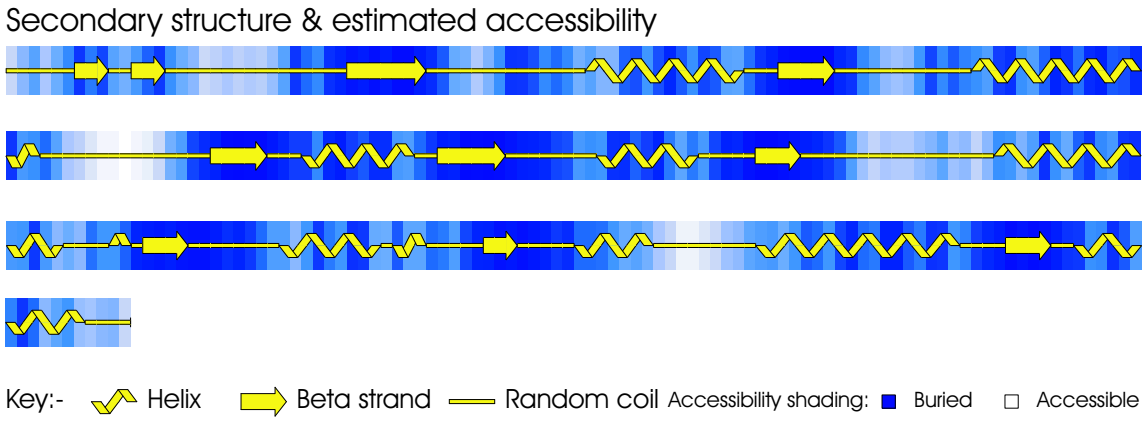


Figure 2.9: Results of the PROCHECK quality check of the PfDHPS model (only selected results are shown). Top: image represents the Ramachandran plot. The three residues in the disallowed regions are a result of the modelling process. During subsequent minimization these residue parameters are adjusted to normal. The PROCHECK results of the temlates showed no residues in disallowed regions. Bottom: An estimated solvent acessibility plot. The plot is an estimate of the solvent exposure of each residue. Each vertical color in the secondary structure plot represents a separate residue (100 residues per plot).

site of PfDHPS. The model indicated that sulfadoxine fills up the remainder of active site cavity (after DHP has docked) and thus blocks the entrance of *p*-aminobenzoic acid into the active site. Sulfadoxine also shows strong interactions with the pyrophosphate group on DHP, which helps to keep it in the active site. The methyl groups on sulfadoxine show hydrophobic interactions with the side chain of Lys 540 in PfDHPS. When Lys 540 is mutated to Glu there is a charge reversal, as well as a shortening of the hydrophobic side chain (Figure 2.10). This may cause a loss of the hydrophobic interaction and thus the methyl groups are forced to move, resulting in a less tight interaction between PfDHPS and sulfadoxine. The A581G simulation showed that the methoxy groups move away from residue 581 due to a decrease in hydrophobic interactions. Ser 436 shows interaction with sulfadoxine and when the S436F/A mutation occurs there is a loss of this interaction. The S436F mutation causes steric hindrance as the Phe mutation takes up the space needed for sulfadoxine to bind. The S436A mutation will cause a loss of interaction as the interaction between Ser 436 and sulfadoxine is through the hydroxyl group of Ser 436.

2.3.3. Homology modeling of PPPK

The template-target alignment was obtained from the multiple sequence alignment and used for modelling. The alignment showed two large insertions (average of 90 residues each) and these inserts were excluded from the model building alignment due to their length as well as the lack of a template match. There is some conservation between the Plasmodial species but the *P. vivax* sequence seems to have diverged the most (Figure 2.11).

Unlike the Plasmodial DHPS sequences, the length of the inserts in the Plasmodial PPPK sequences seems to be more conserved between the Plasmodial species (except for *P. falciparum*). Insert 1 shows a high degree of conservation between the five Plasmodial sequences but insert 2 only shows some conservation between *P. falciparum*, *P. chabaudi* and *P. berghei* and only at the beginning of insert 2. Figure 2.13 shows the results of conserved motif detection in PPPK by MEME. The *Plasmodium* family seems to have undergone huge changes in the PPPK protein as seen by the absence of globally conserved regions. Once again there is evidence for the close phylogenetic relationship between *P. chabaudi, P. berghei* and *P. yoelli yoelli* as seen with motif 15 and 16 only occurring in these three species. *P. yoelli yoelli* seems to contain an extra motif 19 which is similar to that of *C. pneumoniae*. *P. falciparum* also contains an extra motif 17 and 18 when compared to the other members of the *Plasmodium* family.

MODELLER produced 10 models and the models were evaluated based on the criteria in section 2.2.1.4. The substrate analogue 87Y (7,8-dihydro-6-hydroxymethyl-7-methyl-7-pterin) and the Mg^{2+} ions were built in from the template crystal structure and modified to the natural substrate using InsightII (Figure 2.15). The active site of the PfPPPK model compared well with the other known crystal structure active sites (Figure 2.14), with differing residues indicated in Table 2.2. The quality of the model and the templates were checked by PROCHECK and the results for the model is shown in

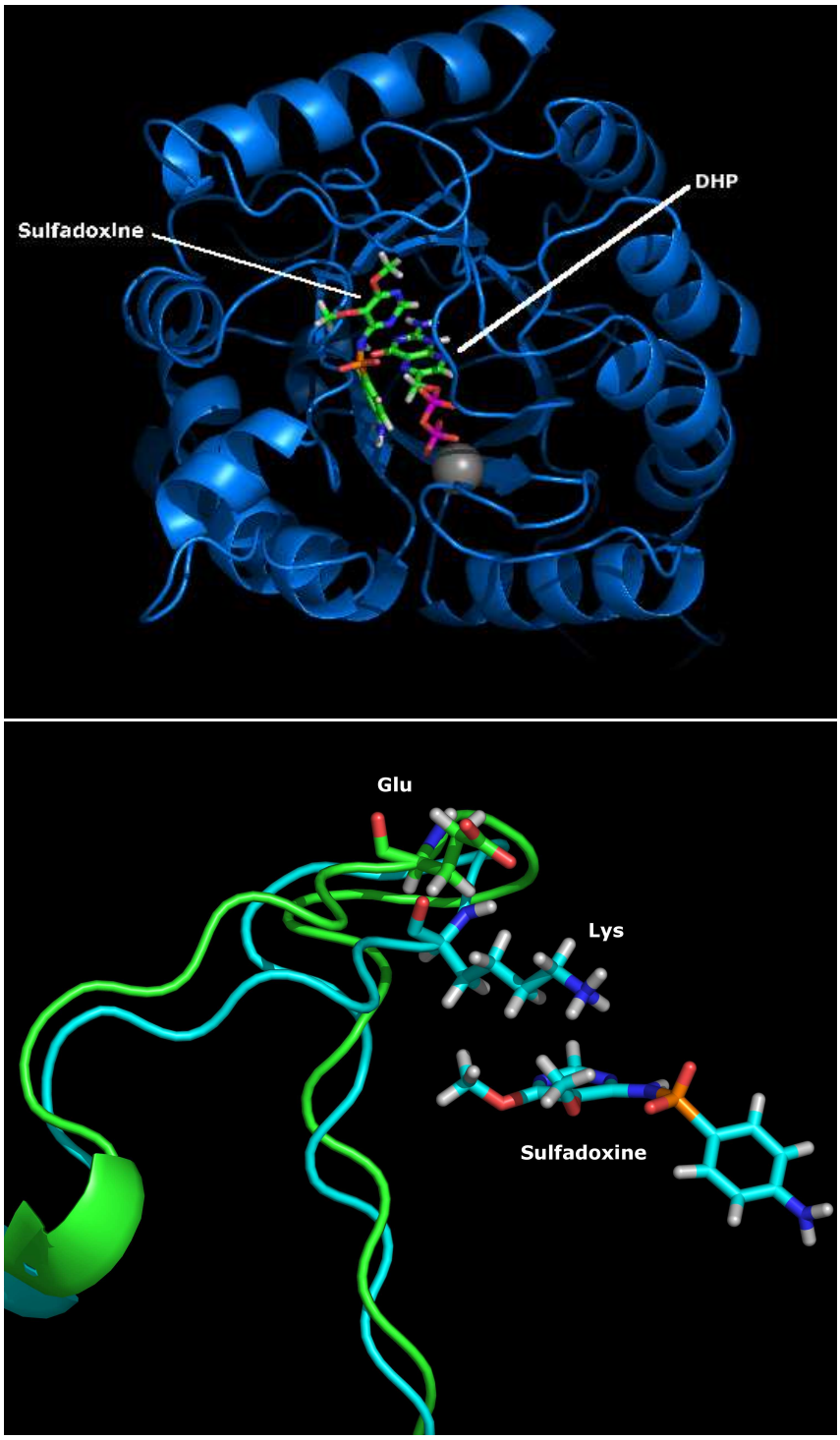


Figure 2.10: Top: The interaction between PfDHPS, DHP and sulfadoxine. The grey sphere is the Mg^{2+} ion, which anchors DHP through the pyrophosphate groups. The *p*-aminobenzoic acid analogue part of sulfadoxine assumes a marginally different orientation than *p*-amino benzoic acid in Figure 2.8 and shows interactions with the pyrophosphate group on DHP. Bottom: The effect of the K540E mutation on sulfadoxine binding. In the wild-type (light blue) the methoxy groups on sulfadoxine shows interactions with the sidechain of Lys 540. When the K540E (green) mutation occurs, the methoxy groups rotate to point away (not shown) and the distance between residue 540 (Glu) and sulfadoxine doubles. This may be due to the increased interaction of Glu 540 with the solvent, which may also explain why the loop opens up more than in the wild-type. Generated with PYMOL (DeLano 2002, http://www.pymol.org).

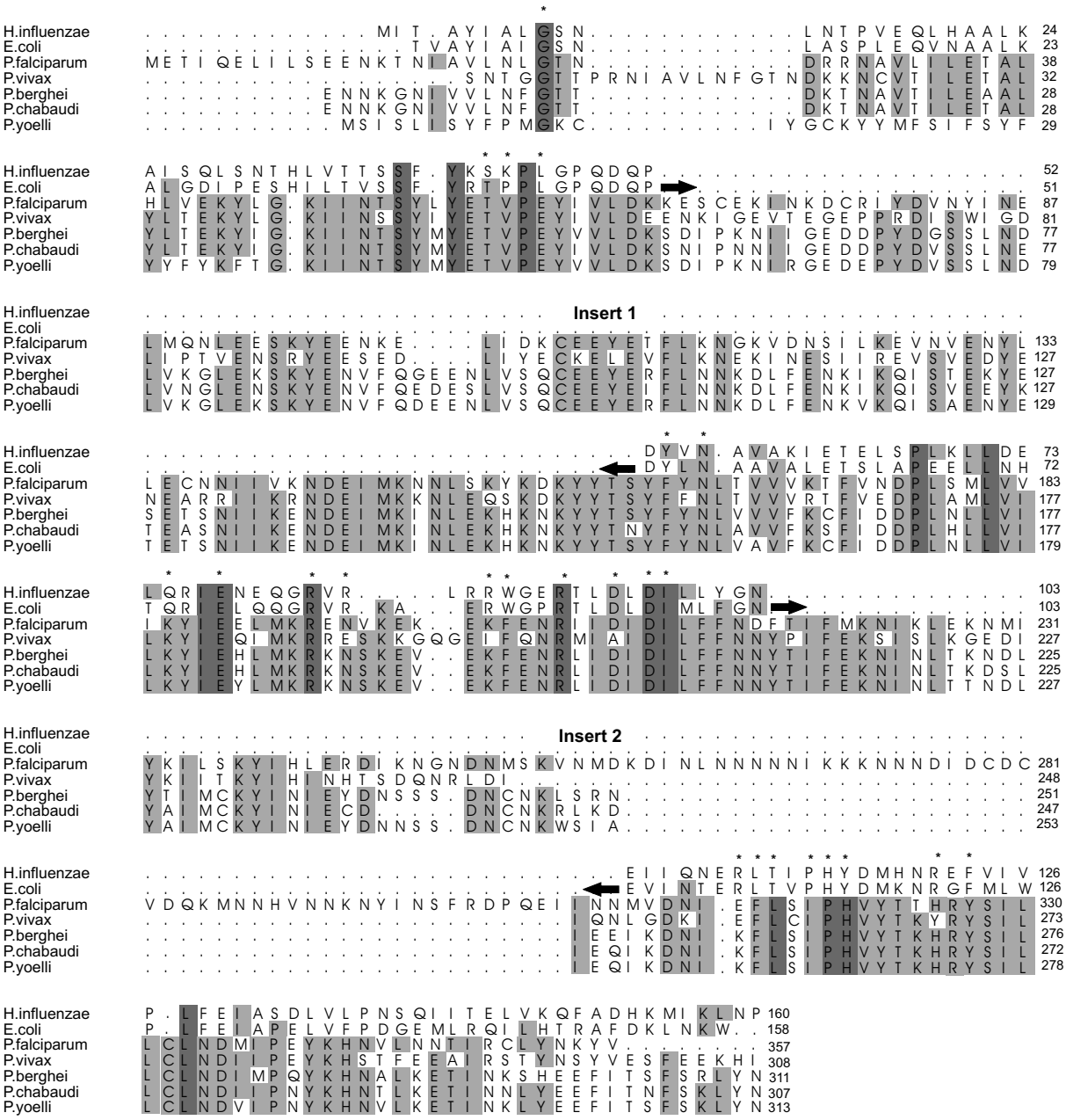
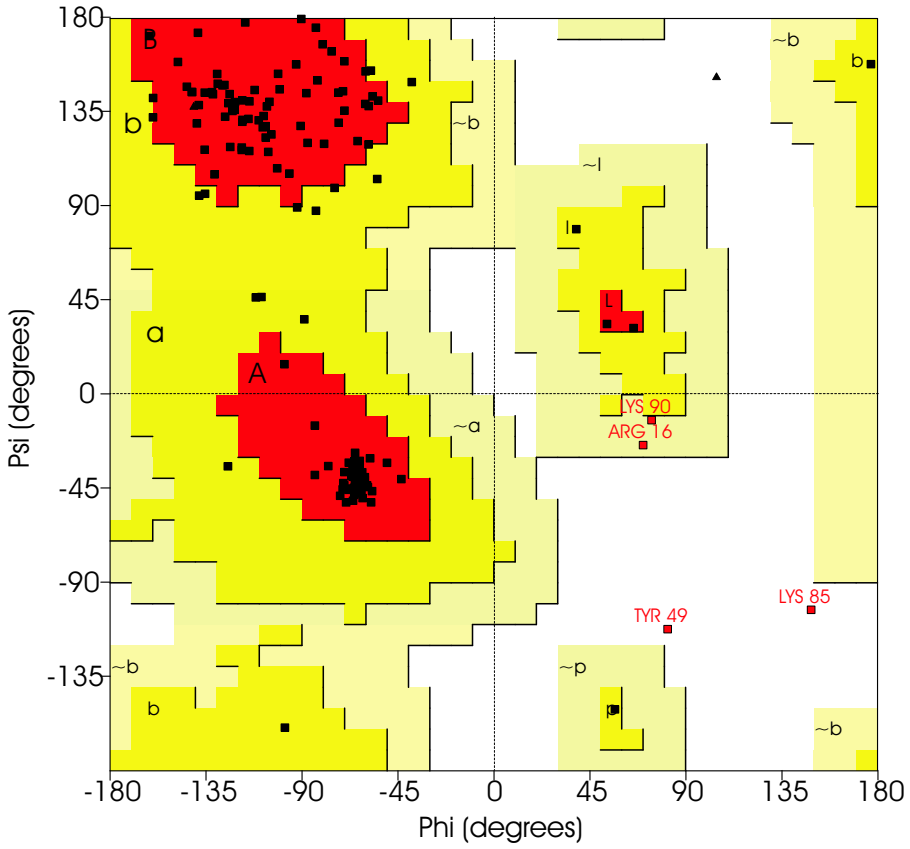


Figure 2.11: The multiple sequence alignment for the complete *E. coli*, *H. influenzae* PPPK and five Plasmodial PPPK sequences. Notice the two large inserts in the *Plasmodium* family when compared to *E. coli* and *H. influenzae*. Insert 1 shows some conservation between the Plasmodial species but insert 2 only has a short conserved motif at the beginning. There are also a few short insertions as seen in insert 1 for *P. berghei*, *P.chabaudi* and *P. yoelli yoelli*. The large inserts (insert 1 and 2) were left out of the modelling process due to their excessive length and the lack of a template match. '*' indicates PPPK contacts with ATP and the substrate analogue. Note the additional *P. falciparum*-specific insert in insert 2. This specific insert is a low-complexity region. PfPPPK numbering stops at 357 and PfDHPS numbering starts at 359, this includes a short linker region between PPPK and DHPS.



Plot statistics

Residues in most favoured regions [A,B,L]	129	85.4%
Residues in additional allowed regions [a,b,l,p]	18	11.9%
Residues in generously allowed regions [~a,~b,~l,~p]	2	1.3%
Residues in disallowed regions	2	1.3%
Number of non-glycine and non-proline residues	151	100.0%
Number of end-residues (excl. Gly and Pro)	3	
Number of glycine residues (shown as triangles)	2	
Number of proline residues	4	
Total number of residues	160	

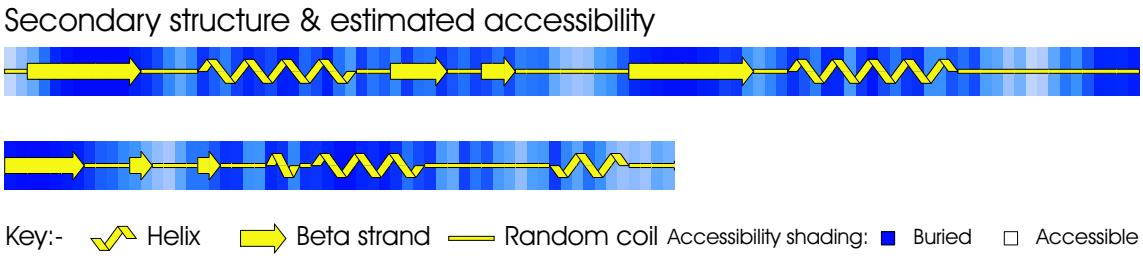


Figure 2.12: Results of the PROCHECK quality check of the PfPPPK model (only selected results are shown). Top: Ramachandran plot of PfPPPK. The two residues in the disallowed regions are due to the modelling process. The residue parameters are shifted to normal during subsequent minimizations. Bottom: An estimated solvent acessibility plot. The plot estimates the solvent exposure of each residue. Each vertical color in the secondary structure plot represents a separate residue (100 residues per plot).

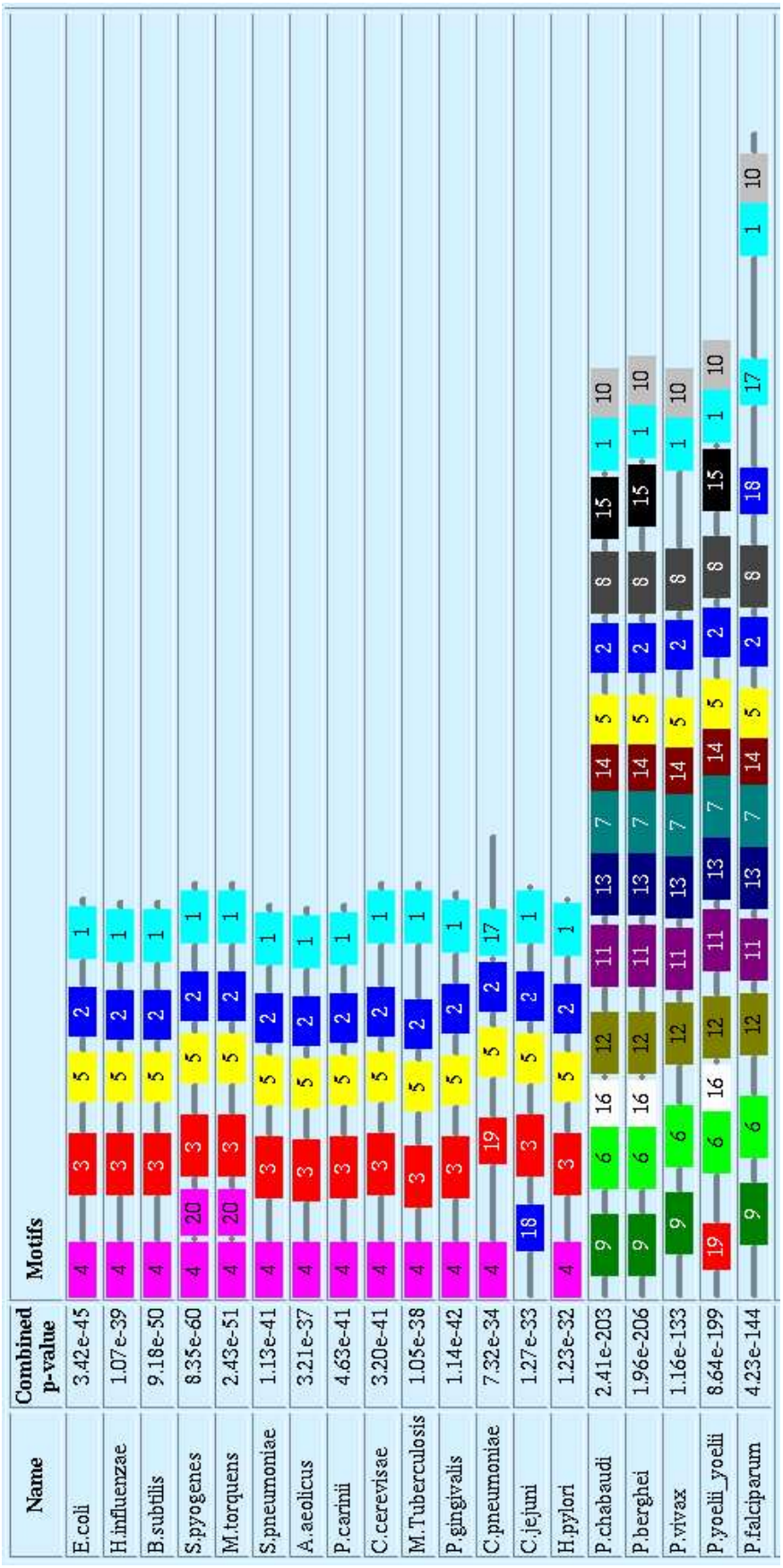
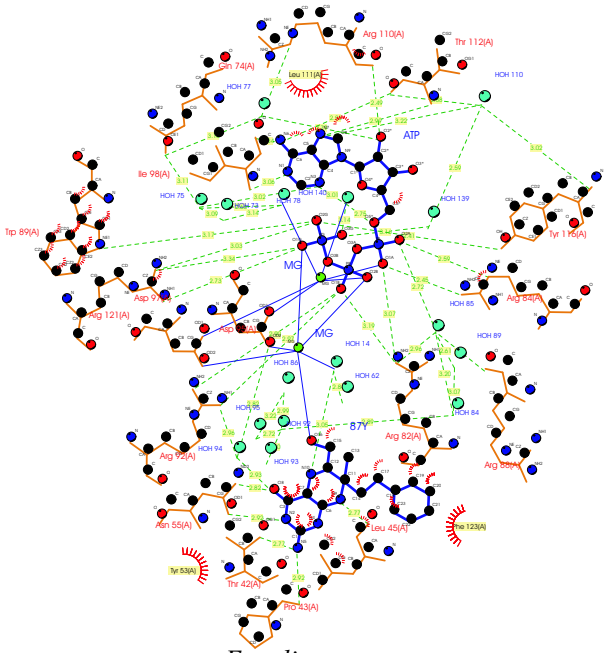


Figure 2.13: The conserved motifs in PPPK as detected by MEME. Note the large inserts in the *Plasmodium* species when compared to the other species.

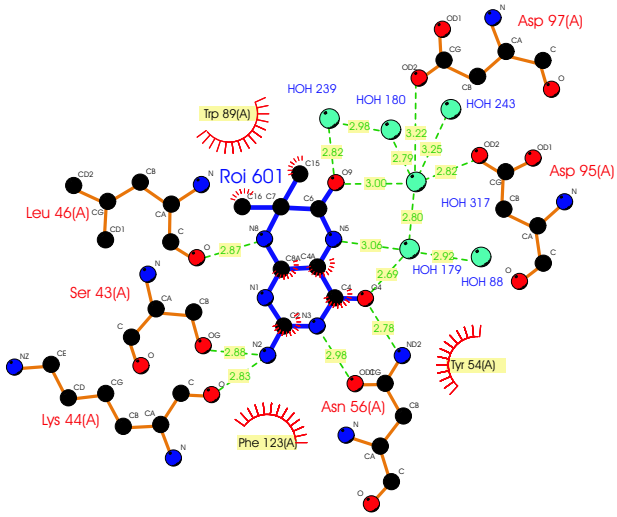
Figure 2.12. During minimization the Mg^{2+} ions anchor the phosphate groups on ATP and prevents it from moving too far away. ATP also remained close enough to the pyrophosphate- accepting group for the reaction to occur. From the crystal structure of *E. coli* it can also be seen that the Mg^{2+} ions play a role in anchoring the phosphate groups of ATP as well as mediating interactions between the protein and the substrate (Figure 2.14).

Table 2.2 shows the interactions between PPPK, substrate and ATP. Blaszczyk *et al.* (2000) remarked that the differences seen in substrate-interacting residues may be responsible for the different isozymes of PPPK isolated from biological sources. These authors also proposed that the following residues are involved in loop stabilization in *E.coli* PPPK: Asn 10, Gln 50, Arg 82, Arg 84, Arg 88, Trp 89, Arg 92 and Tyr 116 (Pf: Asn 25, Asp 65, Asn 100, Glu 105, Lys 106 and His 134). The movement of the loops will be discussed in Chapter 3. PPPK contains two Mg^{2+} ions which are both six-coordinated. Mg1 interacts with the α-phosphate and β-phosphate groups while Mg2 interacts with β-phosphate and the γ-phosphate group. The Mg^{2+} ions also interact with the oxygen atoms in the sidechains of Asp 95 and Asp 97 (Pf Asp 113, Asp 115). Both the Mg^{2+} ions also coordinate with water molecules. Blaszczyk *et al.* (2000) noted that Asp 95 and Asp 97 have to undergo significant conformational changes to interact with Mg^{2+} and thus magnesium binding-induced fit is evident.

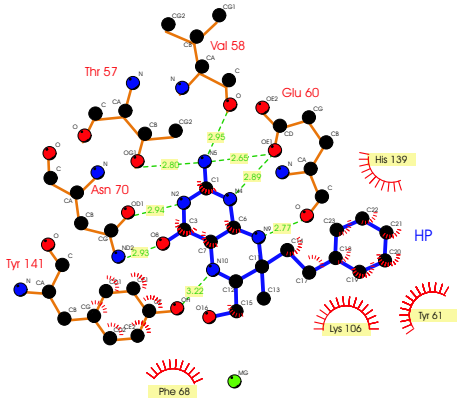
The *E.coli* structures also contained a substrate analogue (87Y) as well as the natural substrate HP (6-hydroxymethyl-7,8-dihydropterin). In the *E. coli* crystal structure HP undergoes ring stacking to stabilize the double ring system (Figure 2.16). Tyr 53 and Phe 123 are involved in ring stacking in the *E. coli* structure but in the PfPPPK model the situation appears to be reversed with Phe 68 fulfilling the role of Tyr 53, and Tyr 141 replacing Phe 123. HP also shows hydrogen bonds with *E. coli* residues Thr 42, Pro 43, Leu 45, Asn 55 (Pf Thr 57, Val 58, Glu 60, Asn 70) and these interactions saturate the hydrogen bond capacity of HP. Five other residues are also involved in binding HP: *E. coli* Gly 8, Tyr 53, Trp 89, Asp 95 and Phe 123 (Pf Gly 23, Phe 68, Lys 106, Asp 113, Tyr 141). Gly 8 is highly conserved between the different species and cannot be replaced by any other residue as this would interfere with HP binding through steric constraints, as well as preventing Asn 55 from changing its conformation to allow HP binding.

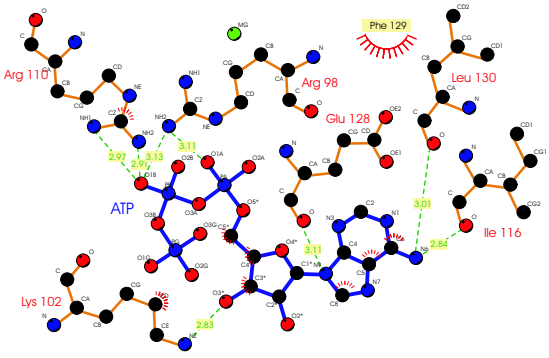


E. coli



H. influenza





P. falciparum

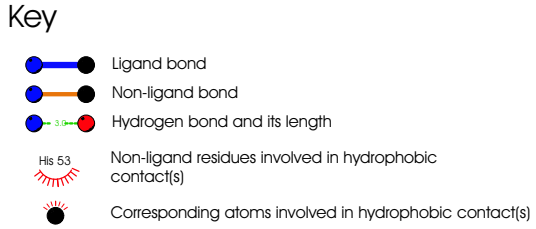


Figure 2.14: A comparison between the active site interactions of the *P. falciparum* PPPK model and known PPPK crystal structures. 87Y (*E. coli*) and Roi (*H. influenzae*) are the substrate analogues in the crystal structures. HP is the substrate analogue from the crystal structures built into the *P. falciparum* PPPK model. All figures were generated with LIGPLOT (Wallace *et al.*, 1995).

Table 2.2: Active site residue comparison between the crystal structures of *E. coli*, *H. influenzae* PPPK and the *P. falciparum* PPPK model. Some residues differ but the type of interaction between the residue and the substrate analogue remains the same. Note that most of the ATP-interacting residues interact with ATP though amide and/or carbonyl groups and thus the sidechain nature of these residues are relatively unimportant. Adapted from Blaszczyk *et al.* (2000).

E. coli	***H. influenzae***	***P. falciparum****	**Interaction with PPPK**
Gly 8	Gly 9	Gly 11/23	Interacts with HP, highly conserved
Thr 42	Ser 43	Thr 45/57	Forms hydrogen bond with primary amine on HP
Pro 43	Lys 44	Val 46/58	Use peptide backbone O for HP interaction
Leu 45	Leu 46	Glu 48/60	Glu 60 uses εO to form hydrogen bond with HP, Leu 45 uses backbone O for hydrogen bond formation
Tyr 53	Tyr 54	Phe 68/161	Sidechain associates with HP, substrate sandwiched between Tyr (Ec 53, Hi 54) and Phe (Ec F123, Hi F123)
Asn 55	Asn 56	Asn 70/163	Interacts with the 4-oxo group on ATP
Gln 74	Gln 76	Lys 78/185	Interacts with ATP through the NH_X groups of Lys and Gln
Glu 77	Glu 78	Glu 81/188	Interacts with Mg^{2+} and with ATP
Arg 82	Arg 83	Arg 86/193	Interacts with oxygens on the α-phosphate and the β-phosphate group of ATP
Arg 84	Arg 85	Asn 88/195	Interacts with oxygens on the α-phosphate group of ATP, helps with loop stabilization
Arg 88	Arg 88	Lys 93/200	Interacts with ATP through its ϵ-amino group
Trp 89	Trp 89	Lys 94/201	Interacts with substrate and γ-phosphate oxygen of ATP
Arg 92	Arg 92	Arg 98/205	Helps with loop stabilization and interacts with oxygens on the α-phosphate and the β-phosphate group of ATP
Asn 95	Asn 95	Asp 101/208	Interacts with HP and Mg^{2+} through its carboxy oxygens
Asn 97	Asn 97	Asp 103/210	Interacts with Mg^{2+} through its carboxy oxygens
Ile 98	Ile 98	Ile 104/211	Residues interact with ATP through their amide and/or carbonyl groups
Arg 110	Arg 110	Glu 116/313	Arg 110 use backbone O for hydrogen bond with the ribose part of ATP, Glu 128 uses terminal εO
Thr 112	Thr 112	Leu 118/315	Both residues uses peptide backbone O for interaction
Pro 114	Pro 114	Ile 120/317	Forms cavity for ATP, interacts with ATP core
His 115	His 114	Pro 121/318	Forms cavity for ATP, interacts with ATP core
Tyr 116	Tyr 116	His 122/319	Forms cavity for ATP, interacts with ATP core
Arg 121	Arg 121	His 139/324	Forms cavity for ATP, interacts with ATP core
Phe 123	Phe 123	Tyr 141/326	Sidechain associates with HP, substrate sandwiched between Tyr (Ec 53, Hi 54) and Phe (Ec F123, Hi F123)

* The first number indicates the residue in the model. The second number indicates the residue in the full bifunctional protein.

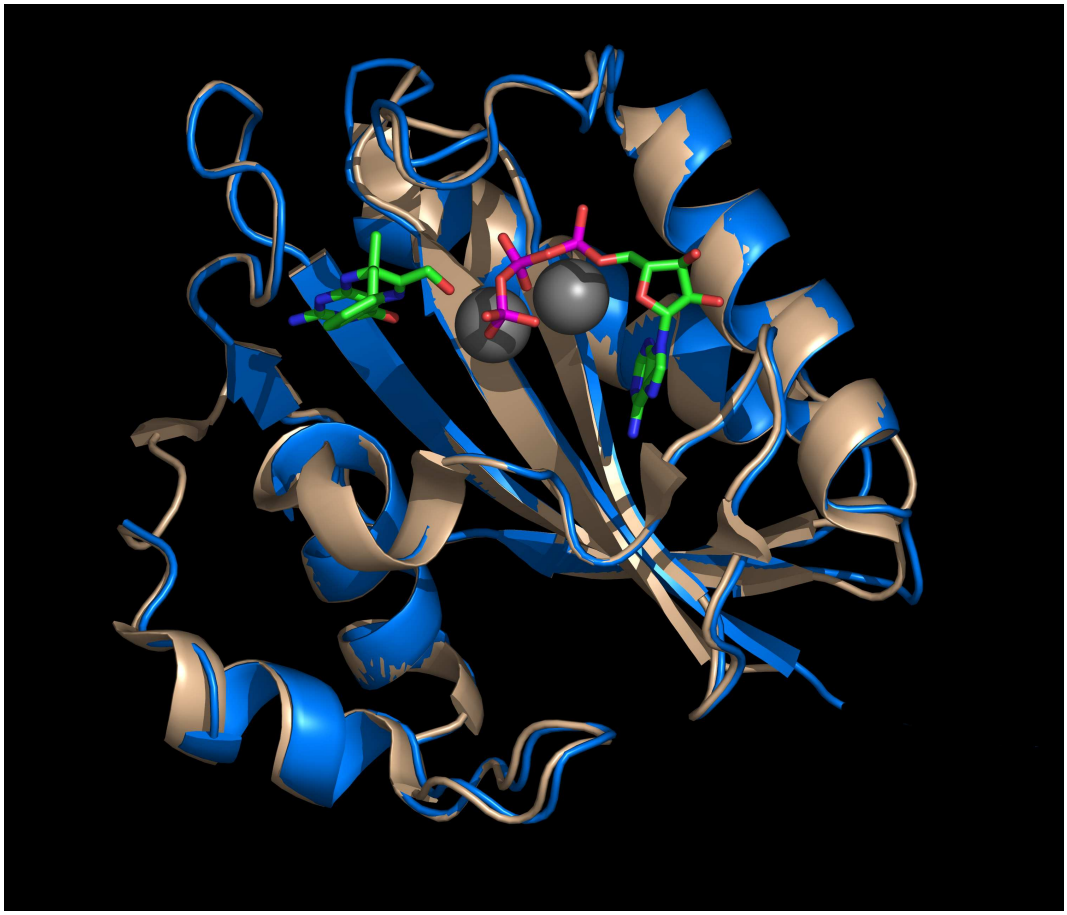


Figure 2.15: A comparison of *E. coli* PPPK (beige) with the PfPPPK model (blue). The model does not contain any of the parasite-specific insertions. The substrate analogue, 87Y is the stick model on the left and ATP is the stick model on the right. The Mg^{2+} ions are grey spheres. Note how the Mg^{2+} ions anchor the ATP. All the interactions are shown in Table 2.2. Generated with PYMOL (DeLano 2002, http://www.pymol.org).

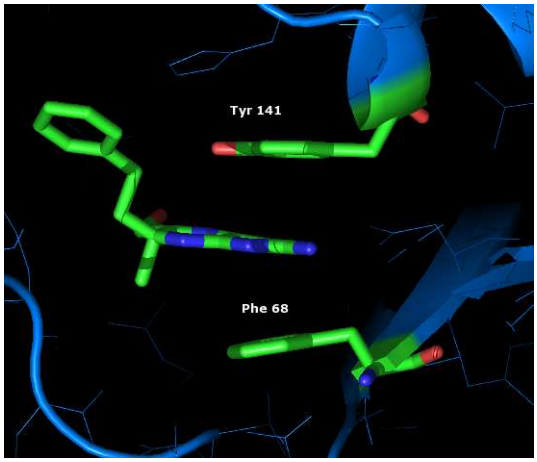


Figure 2.16: The ring stacking used to keep the substrate analogue in place in PPPK. *E. coli* uses Tyr 53 and Phe 123 while in the PfPPPK model the situation is reversed with Phe 68 and Tyr 141 being used. Generated with PYMOL (DeLano 2002, http://www.pymol.org).

2.4. Discussion

2.4.1. DHPS Homology Model

The DHPS model will be discussed in three sections, with the first section focusing on the homology model, the second section on the parasite-specific inserts and their effect on the PfDHPS model and the third section on ligand-protein interactions.

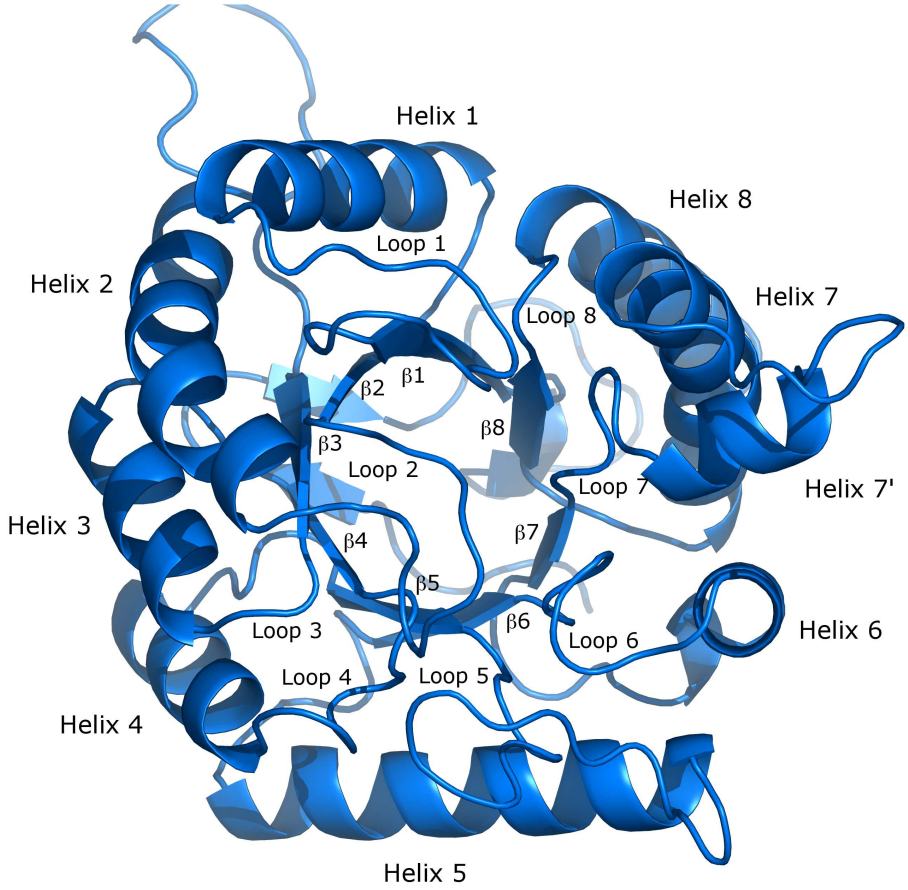


Figure 2.17: The modelled structure of *P. falciparum* DHPS. The model maintains the core barrel structure of *M. tuberculosis* (9 α-helices and 8 β-sheets). Helices 6, 7, 7' and 8 have been proposed to be involved in dimer interaction in *M. tuberculosis* and *S. aureus*, however only helix 7 and 8 are involved in dimer formation in *B. anthracis* (Babaoglu *et al.*, 2004). Insert 2 is located just after helix 7'.

2.4.1.1. *P. falciparum* DHPS model

The *P. falciparum* DHPS homology model compared well to the available crystal structures, especially in the active site region. The major structural difference between the model and the crystal structures is the inserts in the *P. falciparum* model. The annotated structure of *P. falciparum* DHPS is shown in Figure 2.17. The longest insert (insert 2, 39 residues) was left out in the construction of the model. No template is available and this region would thus only form an unordered peptide chain. In the crystal

structure of *B. anthracis* and *E. coli*, there is an anti-parallel β-sheet at the bottom of the TIM-barrel but not in the crystal structures of *S. aureus* and *M. tuberculosis*. The model contains an anti-parallel β-sheet at the bottom of the barrel and this is based on the *B. anthracis* template used. Support for or against the existence of this region as an anti-parallel β-sheet in *P. falciparum* DHPS could not be found. An interesting feature of DHPS structures in general is that the loops of the N-terminal end of the barrel sheets between the sheets and helices are very short (+/- 7 residues) while the C-terminal end has very long loop sections (some in excess of 30 residues). This correlates with the observation that the active site is located at the C-terminal end of the barrel sheets in TIM-barrel enzymes (Nagano *et al.*, 2002). Baca *et al.* (2000) and Hampele *et al.* (1997) speculated that loops 1, 2, 5 and 7 play a role in catalysis and that helices 6, 7, 7' and 8 play a role in protein-protein interactions in *M. tuberculosis*. The same loops are implicated in catalysis in *B. anthracis* but protein-protein interactions are mediated only by helices 7 and 8 (Babaoglu *et al.*, 2004).

2.4.1.2. *P. falciparum* DHPS-Specific Inserts

As seen from the alignment (Figure 2.4) *P. falciparum* DHPS contains two inserts in comparison to the template used for modelling. The high conservation of insert 2 between five *Plasmodium* species, implies the need to retain this part of the insert. The functional importance of inserts has been proven for *P. falciparum* dihydrofolate reductase-thymidilate synthase (DHFR-TS, Yuvaniyama *et al.*, 2003) and S-adenosylmethionine decarboxylase/ornithine decarboxylase (AdoMetDC/ODC, Birkholtz *et al.*, 2004). Yuvaniyama and co-workers have shown that the insert between the DHFR and the TS domains (containing two short α-helices) are involved in dimerization and domain organization. The helix in the insert serves as an anchor between the domains thus helping to keep the domains together. Birkholtz *et al.* (2004) deleted various combinations of parasite-specific inserts to determine their functionality. When certain inserts were deleted a 95% loss in enzyme activity was observed. When the connecting hinge region between the two domains were deleted an activity loss of 33% for AdoMetDC and a $\approx$50% loss in ODC activity was observed. When the domains were co-expressed with their respective inserts deleted, they also observed a marked loss in activity as well as a decrease in the formation of a functional, hetero-tetrameric enzyme. These results suggest the importance of inserts in domain-domain interactions as well as in enzyme activity.

The inserts in the known Plasmodial DHPS sequences differ with respect to their length and conservation. Insert 1 showed a difference in insert length between the species with *P. chabaudi* being the longest and *P. falciparum* the shortest. *P. chabaudi*, *P. berghei* and *P. yoelli yoelli* sequences show a high degree of conservation for insert 1 but the reason why it is not conserved between all five species is not known. A possible reason for this may be that these three species are more closely related to each other than

they are to *P. vivax* or *P. falciparum*. Kedzierski and co-workers (2002) used the adenylosuccinate lyase gene to do a phylogenetic analysis of the *Plasmodium* family. In this study they came to the conclusion that there are three main branches in the *Plasmodium* family. *P. berghei, P yoelli yoelli* and *P. chabaudi* were closely related on one branch while *P. vivax* and *P. falciparum* each occupied a separate branch. This supports the observation that the highly conserved sequences in insert 1 of *P. berghei, P. yoelli yoelli* and *P. chabaudi* are due to a recent common ancestor in the evolution of the *Plasmodium* family (Kedzierski *et al.*, 2002).

Insert 2 is interesting with respect to the secondary structure predicted for the last residues (747-759) in the *P. falciparum* sequence. The fact that the predicted helix is highly conserved between all five species may suggest that it has a functional role in the Plasmodial PPPK-DHPS enzyme. As seen from the work of Yuvaniyama *et al.* (2003), it may be involved in domain-domain interactions either between the PPPK-DHPS domains or between similar domains (DHPS-DHPS or PPPK-PPPK) and may also influence enzyme activity. Once again (as for insert 1) the highest sequence conservation for insert 2 is between *P. berghei, P. yoelli yoelli* and *P. chabaudi* but in this insert (in contrast with insert 1) *P. falciparum* and *P. vivax* show quite high sequence similarity with the other three species, indicating a need to conserve the sequence of insert 2. Interestingly, *P. vivax* and *P. yoelli yoelli* each contain species-specific inserts with the *P. yoelli yoelli* insert disrupting the predicted α-helix. Another interesting feature is the occurrence of a repetitive (G/A)KLTNG(D/E) ([Gly,Ala]-Lys-Leu-Thr-Asn-Gly-[Glu,Asp]) motif in the *P. vivax* specific insert. The [G,A]KLTNG[D,E] motif repeats nine times to form the *P. vivax* specific insert. The *P. yoelli yoelli* specific insert displays no such motif, and no repetitive motif was observed in any of the DHPS sequences from other *Plasmodium* species. The function of the repetitive elements remains unknown.

2.4.1.3. Ligand Interactions

The crystal structures of DHPS from *M. tuberculosis* (containing DHP-monophosphate), *B. anthracis* (various models containing pteroic acid, DHP-monophosphate, DHP and MANIC (6-methylamino-5-nitro-iso cytosine)), *E. coli* (sulfate ion) and *S. aureus* (DHP) revealed the conserved contacts made between the substrates and the protein. Table 2.1 lists the most important conserved residues. The interaction between the oxygen on the pyrazine ring and Lys is conserved between all the crystal structures (Pf 609, Mtb 213, Ba 220, Ec 221, Sa 203) as well as the interaction between the amino group on the pyrazine ring, Asn and Asp (Pf 503 and 575, Mtb 105 and 177, Ba 120 and 184, Ec 115 and 185, Sa 103 and 167). Asp also shows a conserved interaction with N8 of the pyrazine ring (Pf 482, Mtb 86, Ba 101, Ec 96, Sa 84). From the *S. aureus* and *B. anthracis* crystal structures the position and orientation of the pyrophosphate groups of DHP can be seen. The pyrazine ring structure also undergoes $\pi - \pi$-stacking

interactions with Arg (Pf 686, Mtb 253, Ba 254, Ec 255, Sa 239). It is well known that the pyrophosphate group interacts with metal ions. The position of the metal ion was determined from the structures of *S. aureus* (Mn^{2+}) and *M. tuberculosis* (Mg^{2+}). Kasekarn *et al.* (2004) determined that *P. falciparum* uses a Mg^{2+} ion for catalysis while *S. aureus* uses a Mn^{2+} ion. The *B. anthracis* and *M. tuberculosis* structures show that some of the oxygen atoms on the pyrophosphate group interact with the metal ion and with Arg (Pf 686, Mtb 253, Ba 254, Ec 255, Sa 239) as well as with His (Pf 688, Mtb 255, Ba 256, Ec 257, Sa 241). The metal ion is also anchored by the His residue (Pf 688, Mtb 255, Ba 256, Ec 257, Sa 241). This orientation of the pyrophosphate group exposes the carbon atom linking the pyrazine and pyrophosphate group for a SN_2 nucleophilic attack by the amino group of *p*-aminobenzoic acid (Baca *et al.*, 2000, Babaoglu *et al.*, 2004). The pyrophosphate group geometry may also create a partial positive charge on the linking carbon atom.

The structure of *B. anthracis* DHPS provided the only indication of where *p*-aminobenzoic acid is located in the active site. The *E. coli* structure has SAN (a *p*-aminobenzoic acid analogue) in the active site but it is in an unfavourable position to form pteroic acid. The *B. anthracis* structure indicated that the benzene ring of *p*-aminobenzoic acid undergoes hydrophobic interactions with the side chain of Lys (Pf 609, Mtb 213, Ba 220, Ec 221, Sa 203). The conformation of Lys 220 changes to accommodate the hydrophobic interaction with *p*-aminobenzoic acid when comparing the crystal structure with a ligand to the unbound structure. As a result of the *p*-aminobenzoic acid stacking with Lys, the carboxy group of *p*-aminobenzoic acid is in close enough proximity to form hydrogen bonds with the γ-O oxygen and the amide backbone nitrogen of Ser (Pf 607, Mtb 213, Ba 221, Ec 219, Sa 201).

As sulfadoxine is a *p*-aminobenzoic acid analogue it maintains most of the same interactions as *p*-aminobenzoic acid. The carboxyl group interactions of *p*-aminobenzoic acid are however replaced by the two oxygen atoms on the sulfur. This causes *p*-aminobenzoic acid and sulfadoxine to compete for the active site. Sulfadoxine may have some hydrophobic associations with the sidechain of Lys 540 through the methoxy groups. The K540E mutation shortens the sidechain with two methyl groups as well as reversing the charge. This may cause a loss of hydrophobic interaction with sulfadoxine. Sulfadoxine may also have some added interactions with Ala 437 although the orientation of Ala 437 did not indicate this. The S436F/A mutation causes a loss of the possible interaction between Ser 436 and sulfadoxine. The S436F mutation causes steric hindrance which prevents sulfadoxine from entering the active site. Ser 436 interacts with sulfadoxine through it's hydroxyl group, which breaks this interaction. Ala 581 shows polar interactions with the methoxy groups on sulfadoxine. The A581G mutation makes the chain more flexible, which may result in a reduced capacity to form polar interactions with sulfadoxine. The A581G does not seem to have a large effect on sulfadoxine binding and this may explain why it is one of the less common mutations in PfDHPS. The model could not explain the A613T/S mutation effects.

2.4.2. PPPK Homology Model

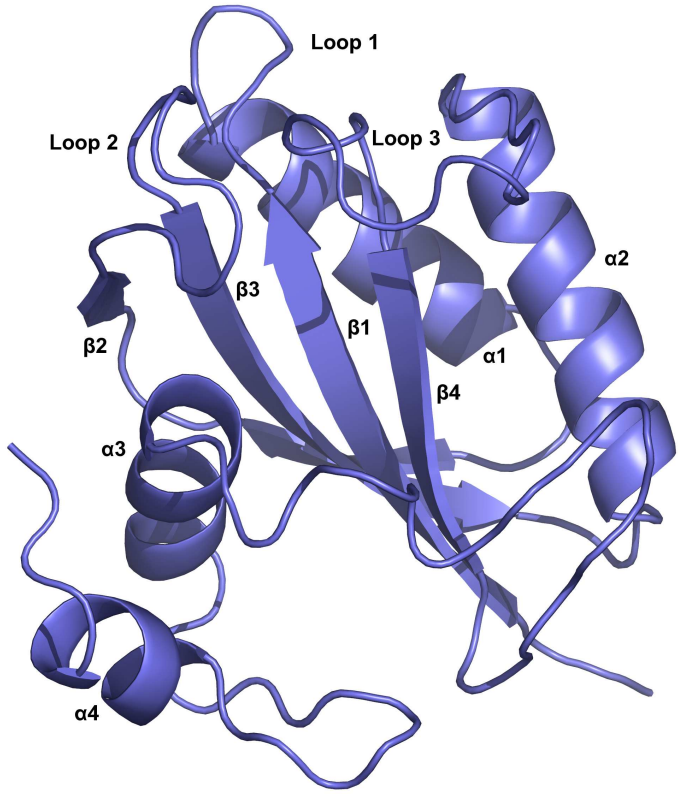


Figure 2.18: The annotated structure of the PfPPPK model. Insert 1 is located after β-sheet 2 and insert 2 is located just before β-sheet 4.

Figure 2.18 shows the annotated PfPPPK model with the loops, helices and sheets indicated (based on the annotation by Blaszczyk *et al.* (2000)). The movement of the loops will be discussed in Chapter 3.

2.4.2.1. PfPPPK inserts

The preliminary alignment showed that some residues aligned between the target and template, but when the model was built the residues did not occupy equivalent coordinates and could thus not fulfill their biological role. This required some of the residues in the alignment to be shifted to ensure that the model represented the active site accurately. When the sequences of the five *Plasmodium* species were compared with those of *E.coli* and *H. influenzae*, the presence of inserts in the *Plasmodium* sequences became apparent. Once again (as is the case with PfDHPS) it seems that *P. falciparum* PPPK contains two inserts in relation to other known species. The alignment (Figure 2.11) supports the hypothesis proposed for DHPS that *P. berghei*, *P. chabaudi* and *P. yoelli yoelli* are more closely related (Kedzierski *et al.*, 2002). From the alignment it is apparent that the second insert is much more conserved between *P. berghei*, *P. chabaudi* and *P. yoelli yoelli* than when compared to those of *P. falciparum* and *P. vivax*.

The three related Plasmodia also share a short insert in the middle of insert 1. In the *P. yoelli yoelli* sequence the first 30 residues do not match up to the other *Plamodium* species. *P. vivax* also seems to contain a short insert at the start of the PPPK sequence. The *P. falciparum* PPPK sequence contains an additional insert inside global insert 2. This is a low-complexity insert as indicated by SEG (Wootton *et al.*, 1993) and is a region rich in Asn residues.

2.4.2.2. Ligand interactions

PPPK contains two vital Mg^{2+} ions in the active site. Blaszczyk *et al.* (2000) suggests six possible functions of the Mg^{2+} ions in PPPK. Firstly, the ions may bring all the catalytic residues into the active site. Secondly, the ions orientate the hydroxyl oxygen of 6-hydroxymethyl-7,8-dihydro-pterin, the β-phosphorous and the oxygen between the α-phosphorous and the β-phosphorous of ATP. Thirdly, the ions may activate the β-phosphorous for a nucleophilic attack. Fourth, the ions lower the pKa of the hydroxyl group on 6-hydroxymethyl-7,8-dihydropterin and thus facilitate the reaction. Fifth, the Mg^{2+} ions can stabilize the transition state of the reaction. Sixth, the removal of AMP may be facilitated by the Mg^{2+} ions. PPPK also has various isozymes and Blaszczyk *et al.* (2000) found that Thr 42 was conserved among eight species but had various other amino acids in other isozymes. In this respect PfPPPK and BaPPPK are similar as they both contain a threonine at postion 42 (Table 3.1).

2.5. Conclusion

Worldwide resistance of the malaria parasite to sulfa-drugs is on the increase and new methods need to be investigated to identify new drugs and drug targets. In the previous sections models for *P. falciparum* DHPS and PPPK were presented. From the crystal structures (as well as the model) it was seen that the active sites of PPPK and DHPS must be able to bind the pterin moiety. Stammers *et al.* (1999) investigated the PPPK and DHPS active sites of *E. coli* to look for similar binding cavities. They found that the hydrogen bonding around the edges of the pterin moiety is the same between the two enzymes but that the phosphate-binding sites are different. In *E. coli,* DHPS interacts directly with the pyrophosphate groups through basic residues. In *E. coli* PPPK the pyrophosphate groups interact through cations (Mg^{2+} ions) as well as hydrogen bonds with basic residues. The same principle holds for the PPPK and DHPS models of *P. falciparum.* Using the pterin moiety as a scaffold, a drug could be designed which would inhibit both enzymes as these enzymes are not found in mammals. This double-pronged approach would alleviate some of the problems of drug resistance in DHPS enzymes.

Using molecular dynamics, the next chapter investigates the effect of resistance-causing mutations on the binding of the natural substrate to DHPS as well as on the binding of sulfadoxine to DHPS.