Chapter 3

Homology modelling of putative malaria drug target proteins

3.1 Introduction

For the purposes of inhibitor design, structural models were needed for the DHFR and TIM enzymes. DHFR from *P. falciparum* has as yet not been crystallised, and TIM was crystallised by Velanker *et al.* (1997) only during the course of this study. Thus, other methods for the preparation of the enzyme structural models had to be attempted. Since the first protein structure was solved, people have been speculating about the possibility of predicting protein structures from the amino acid sequences. Homology modelling has become extremely popular together with the significant increase in the amount of available sequences in nucleotide and protein databases. The crystallographic databases have not grown at the same tempo, but many new sequences are similar to existing proteins in the Brookhaven Protein Databank (PDB) and often the structures of new proteins may be predicted in this way. Although there are approximately 500,000 proteins in GenPept, only about 10,000 experimentally determined structures are available in PDB. Of the known protein sequences, 20-50% have segments that are related to one or more known structures. Approximately 900 folds have been defined out of an estimated few thousand. Projects are be-

ing designed to perform homology modelling on genomes from different organisms (Sanchez and Sali, 1998).

Homology modelling comprises of four basic steps (Sali and Blundell, 1993): the first step is the identification of proteins with known 3D structures that are related to the target sequence. The second step is to prepare optimal sequence alignments of the target with the known structure's sequence. The third step is to build a model for the target sequence given its alignment with the template structures. The fourth step is model evaluation. This cycle may then be repeated as necessary. Various systems are currently available and differ mainly in the way the 3D model is calculated. The first group of methods is based on rigid body assembly (Johnson *et al.*, 1994). A model is constructed from core regions, and loops and side-chains are added from related structures. This assembly involves fitting the rigid bodies on the framework. Another family of methods are based on segment matching (Levitt, 1992). Positions of conserved atoms from the template are used to calculate the coordinates of other atoms. This method employs a database of short segments of protein structure, as well as energy or geometric rules. The third group of methods is based on the satisfaction of spatial restraints (Sali and Blundell, 1993). Restraints are calculated from the alignment of the target protein with the template proteins, and satisfaction of these restraints are attempted.

Our study used homology modelling by satisfaction of spatial restraints, specifically the MODELLER 4 package (Sali and Blundell, 1993). This method can use many different types of information about the target sequence making it probably the most effective of the current approaches. All restraint-based methods determine distance and dihedral angle restraints to be applied on the target structure from the alignment with the template structures and stereochemical restraints and then calculate the model by minimising the violations of these restraints. In MODELLER, many distances and dihedral angle restraints are derived, followed by spatial restraints obtained from the statistical analysis of the relationships between various features of protein structure. Databases of family alignment are used to obtain tables quantifying relationships such as $C\alpha$ - $C\alpha$ distances or dihedral angles. The relationships are expressed as conditional probability density distributions and are used directly as spatial restraints. Spatial restraints are thus obtained empirically from a database and not determined theoretically. Next, the homology-derived restraints and energy terms regard-

ing stereochemistry are combined into an objective function. The model is finalised by optimising the objective function in Cartesian space. In the final iterations, molecular dynamics with simulated annealing is used for model refinement. In this way models with a R.M.S lower than 2Å have been obtained. When the target sequence is at least 40% identical to one or more of the templates, a main-chain Root Mean Square (R.M.S.) error as low as 1Å can be achieved for 90% of the residues. When sequence identity is between 30% and 40%, main-chain R.M.S. error rises to approximately 1.5Å for approximately 80% of the residues. Insertions of larger than 8 residues cannot be modelled accurately. Below 40% sequence identity, it can be expected that about 20% of the residues will be misaligned, but it is possible to edit alignments to prevent large errors. The following errors are commonly found in structures prepared by MODELLER: distortion or shifts in correctly aligned regions (loops, helices, strands), side-chain packing errors, distortion or shifts of regions without template equivalents and distortion or shifts in incorrectly aligned regions (Sanchez and Sali, 1997).

Even though comparative modelling is less accurate than high-resolution experimental techniques, it can be helpful in proposing and testing hypotheses. In some cases, high quality homology models have been used for the design of highly efficient inhibitors. Sudbeck *et al.* (1999) used homology models of JAK3 kinase to design specific high affinity inhibitors for possible cancer trials. Homology models of different cytochrome P450 structures have been utilised along with site-directed mutagenesis to elucidate the molecular determinants of substrate specificity (Szklarz and Halpert, 1997). Already in 1991, *Schistosoma mansoni* serine protease homology models were used to elucidate the function of the enzyme as well as design specific peptide inhibitors (Cohen *et al.*, 1991). Homology modelling can also be used to help guide X-ray and NMR refinement processes.

Some malaria protein structures have been prepared by homology modelling. Lactate dehydrogenase (LDH) was modelled by Hewitt *et al.* (1997). The homology structure was tested by protein engineering and found to be of adequate accuracy. They were able to suggest explanations of the unusual properties of malaria LDH compared with all other LDHs. The malaria cysteine proteinase Falcipain was homology modelled by Li *et al.* (1996), followed by the identification of a series of inhibitors for this protein.

In this study either malaria DHFR or TIM was used as target. All

DHFRs or TIMs with available X-ray structures were then used to obtain optimal alignments with the target. Selected templates were then used for the modelling process. The following goals were set:

- Preparation of an optimal alignment for DHFR
- Homology modelling of DHFR
- Preparation of an optimal alignment for TIM
- Homology modelling of TIM

3.2 Methods

3.2.1 Homology modelling of dihydrofolate reductase and triosephosphate isomerase

All entries for dihydrofolate reductases or triosephosphate isomerases were retrieved from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). After filtering for duplicates varying only by point mutation or bound ligand, alignments together with the *P. falciparum* DHFR or TIM sequences were prepared with the ClustalW package (Thompson *et al.*, 1994), and optimised by visual inspection. Large insertions were removed as necessary by visual inspection.

Alignments were converted to modified PIR format for input in the MODELLER 4 package (Sali and Blundell, 1993). The crystal structures from *E. coli* (1DRA) and *L. casei* (1AO8) were used as templates for model building of DHFR, while all available species were used in the case of TIM. The MODELLER routine for full model building was used on a Silicon Graphics Power Indigo2 Extreme workstation (Silicon Graphics, Mountainview).

The MODELLER log file was inspected for model quality information as compared to empirical guidelines determined from known proteins. Further quality assessments were done with the Procheck package (Laskowski *et al.*, 1996), yielding data such as total Ramachandran plots, individual Ramachandran plots, summaries of bond lengths, angles, etc. The homology structures for *P. falciparum* DHFR or TIM were superimposed with that of the templates and deviations calculated using the ProFit package (University College London).

3.3 Results

3.3.1 Homology modelling of dihydrofolate reductase

Amino acid sequence alignments were prepared using the ClustalW package. An initial alignment was done with DHFR sequences from all species in the Brookhaven Database (Figure 3.1).

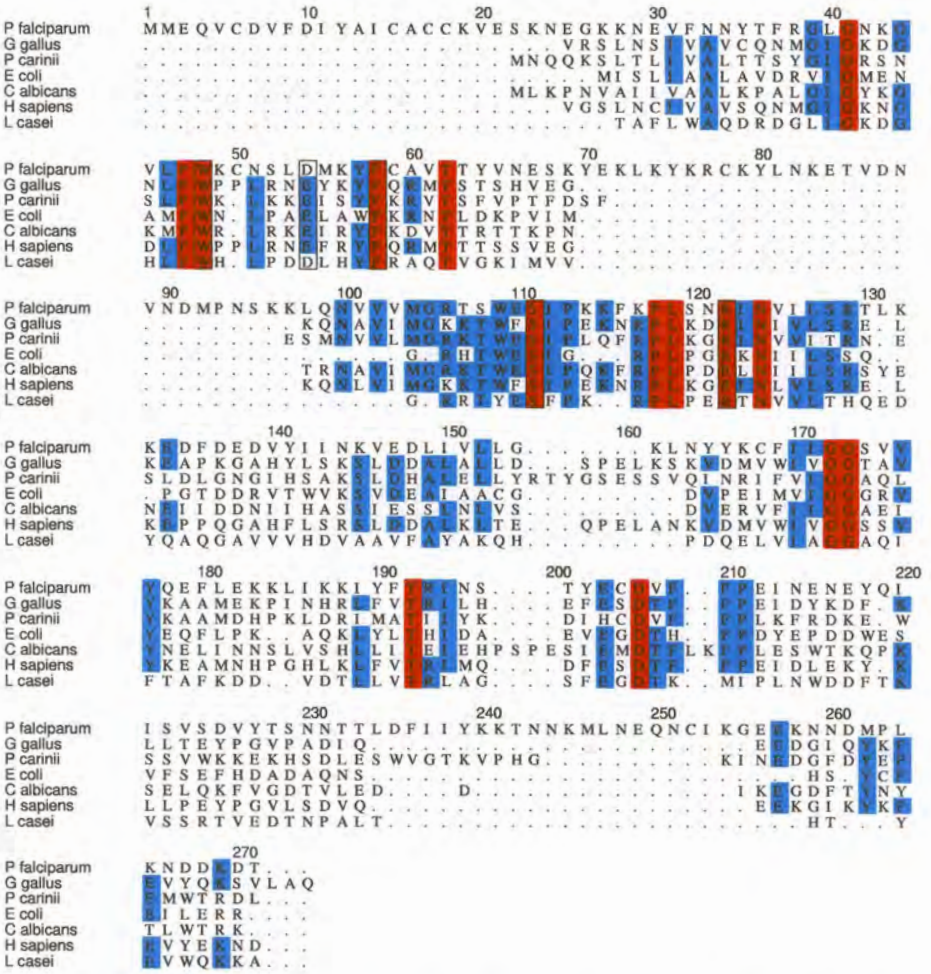


Figure 3.1: Amino acid sequence alignment of malaria DHFR with other species DHFRs from the Brookhaven Protein Data Bank. Homology is indicated in blue and identity in red. Conserved active site residues are boxed.

Malaria DHFR contained a N-terminal segment of 22 amino acids that could not be aligned with any of the other species. Two other large insertions were also present, a 27 residue segment at residue 70-97

and a 38 residue segment at the C-terminal. *P. carinii* DHFR was the only protein also containing an insertion near the C-terminal with another near the middle of the protein. The catalytic residues (*Trp48*, *Glu/Asp54*, *Phe58*, *Ser113*, *Arg122*) were conserved or semi-conserved in all species.

The *E. coli* (1DRA) and *L. casei* (1AO8) DHFR structures were selected as templates being the most homologous to the malaria DHFR. The C-terminal insertion was removed and alignments were reoptimised (Figure 3.2).

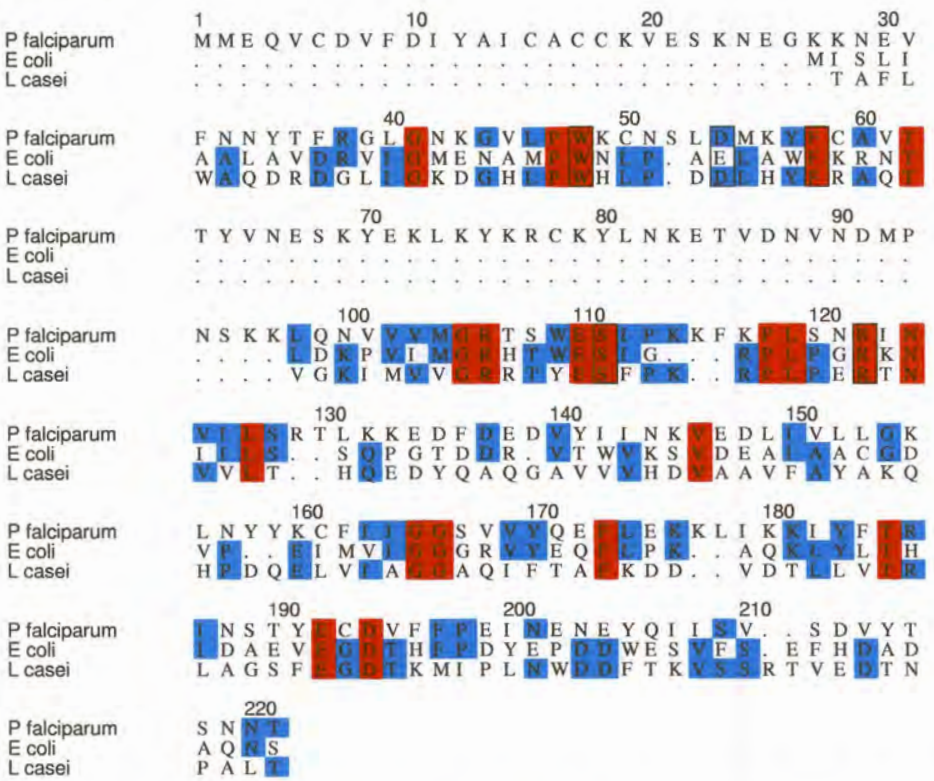


Figure 3.2: Sequence alignment of modified malaria DHFR with *E. coli* and *L. casei* DHFR. Homology is indicated in blue and identity in red. Important catalytic residues are indicated by boxes.

The alignment showed the catalytically important residues to be interrupted by a large insertion of 27 residues. For practical reasons, this insertion together with the N-terminal 22 residues were excised and the alignment re-optimised (Figure 3.3).

The re-optimised alignment was used as input to the MODELLER package. After model building, the output structure was analyzed with the

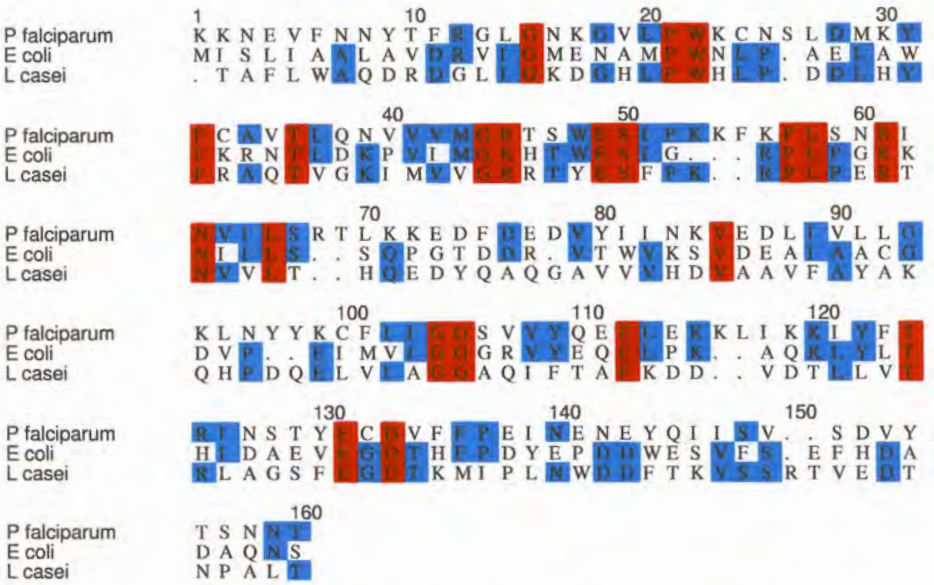


Figure 3.3: Modified sequence alignment of truncated malaria DHFR with *E. coli* and *L. casei* DHFR after removal of large insertions from the malaria sequence. Homology is indicated in blue and identity in red.

Procheck package (not all results shown).

The Ramachandran (Figure 3.4) plot shows 75.2% of residues to be in favoured regions (red), and 16.6% to be in allowed regions (yellow). Of the remaining residues 3.8% were in generously allowed regions (beige) and 4.5% in disallowed regions (white). Based on structures with a resolution of at least 2.0Å, a model of very high quality should show approximately 90% of residues in the favoured regions (Laskowski *et al.*, 1996). None of the catalytic residues were found in problematic regions.

Further analysis of values from the Ramachandran plot showed the main-chain residues to be generally of good quality (Figure 3.5). A bad contact score of 8.8 /100 residues was calculated and an overall G-factor of -0.3. The G-factor is essentially a log-odds score based on the observed distributions of stereochemical parameters, providing a measure of how "normal", or alternatively how "unusual", a given stereochemical property is (Engh and Huber, 1991). When applied to a given residue, a low G-factor indicates that the property corresponds to a low-probability conformation. So, for example, residues falling in the disallowed regions of the Ramachandran plot will have a low (or very negative) G-factor. Similarly for unfavourable chi1-chi2 and chi1

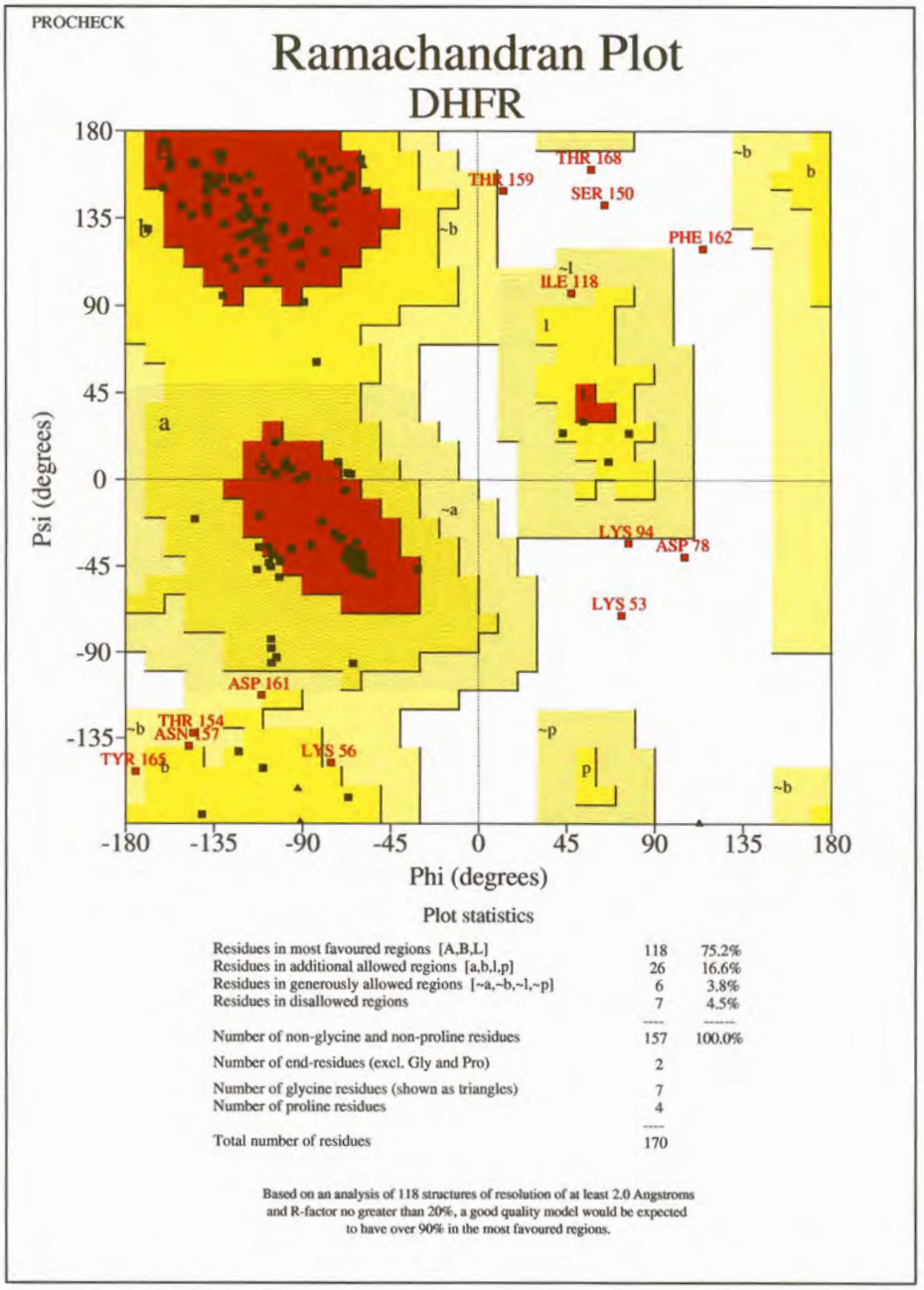
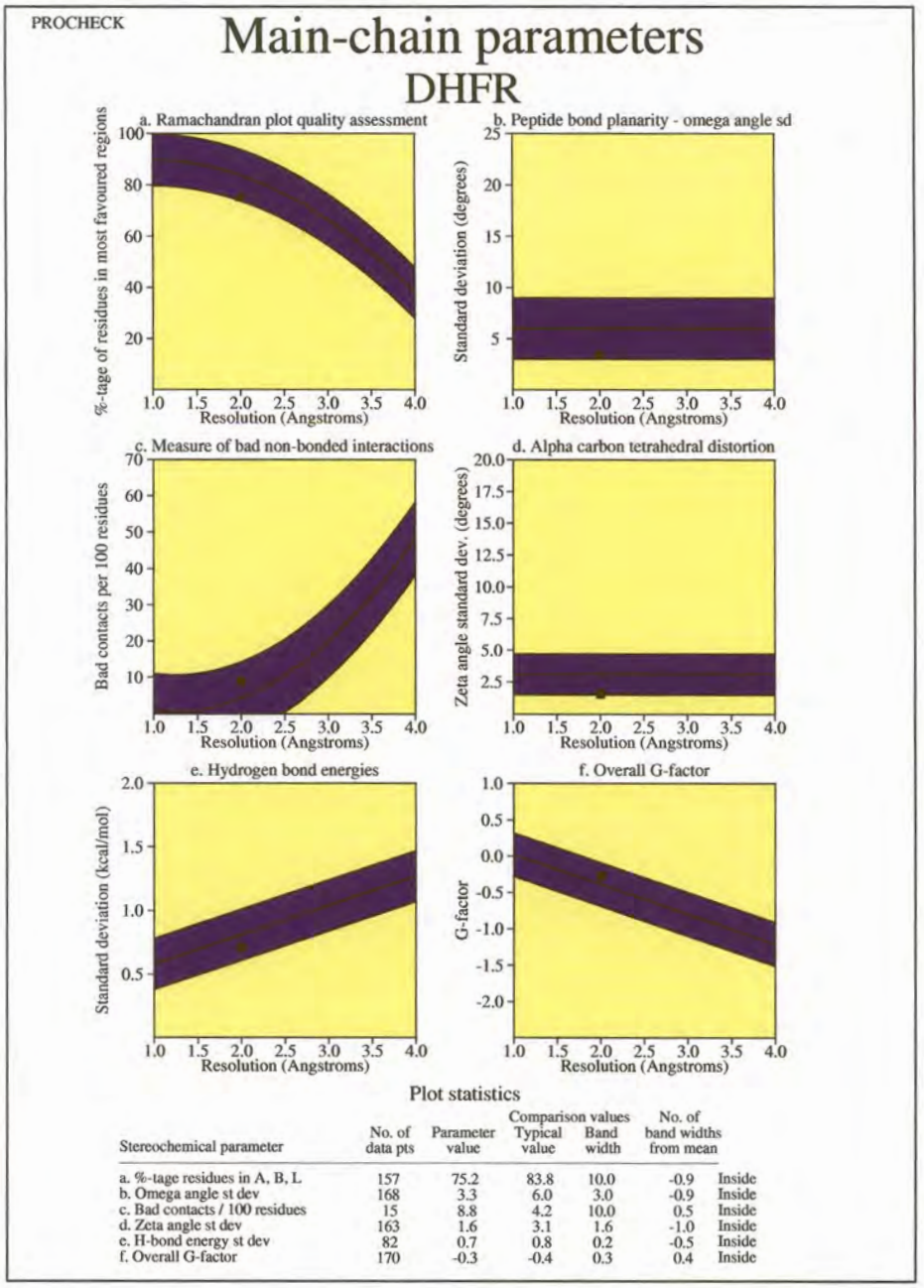


Figure 3.4: Ramachandran plot for the homology model of malaria DHFR.

values.

Side-chain parameters were several deviations away from expected means, and were generally of lesser quality (Figure 3.6). Especially large side-chain chi deviations were calculated.

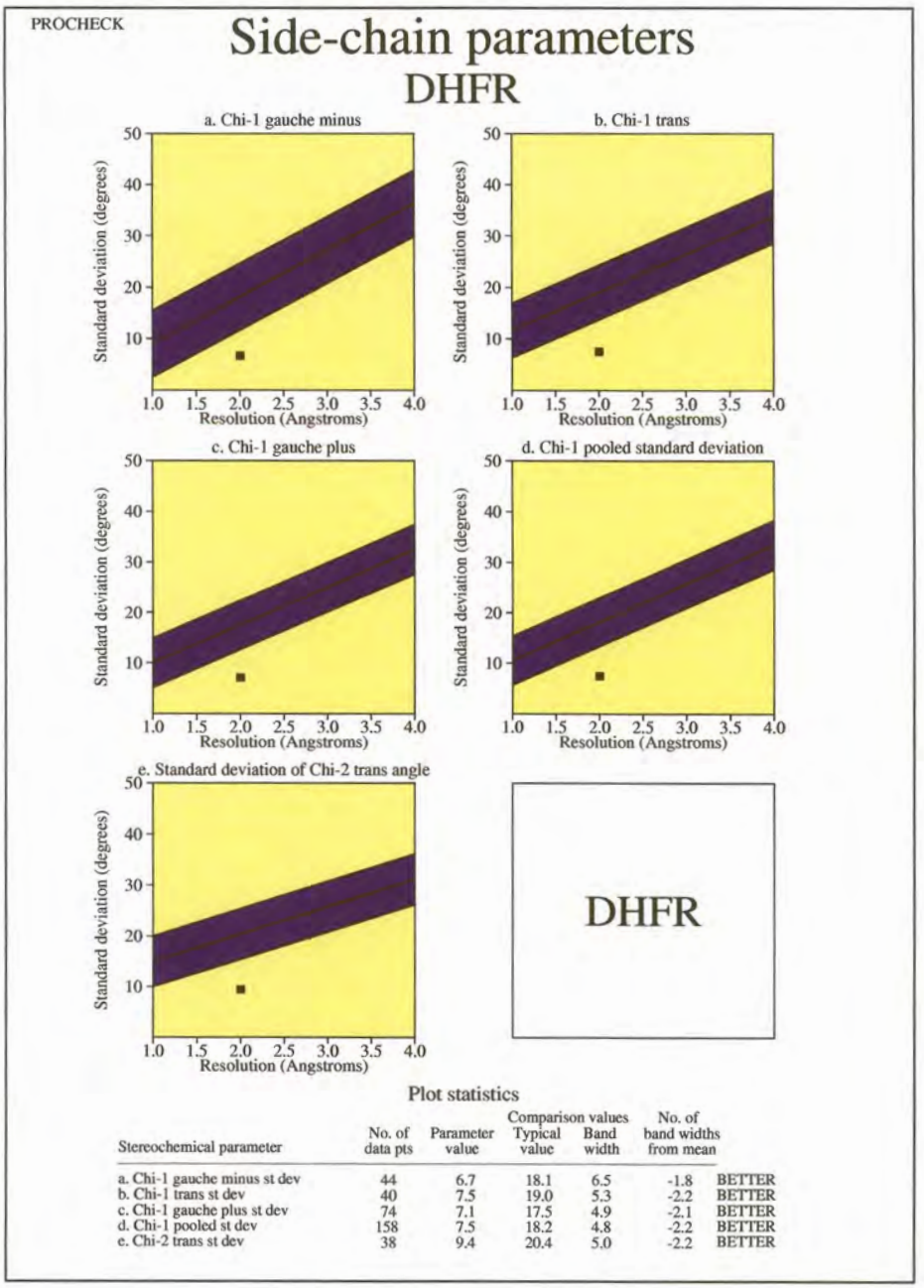


Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	157	75.2	83.8	10.0	-0.9	Inside
b. Omega angle st dev	168	3.3	6.0	3.0	-0.9	Inside
c. Bad contacts / 100 residues	15	8.8	4.2	10.0	0.5	Inside
d. Zeta angle st dev	163	1.6	3.1	1.6	-1.0	Inside
e. H-bond energy st dev	82	0.7	0.8	0.2	-0.5	Inside
f. Overall G-factor	170	-0.3	-0.4	0.3	0.4	Inside

Figure 3.5: Main chain parameter analysis for homology modelled malaria DHFR.

A three-dimensional diagram was prepared indicating the quality values for the model by different colours (Figure 3.7).

A relatively high number of red regions are visible, especially away from the core of the protein and in loop regions. The homology structure was



Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. Chi-1 gauche minus st dev	44	6.7	18.1	6.5	-1.8	BETTER
b. Chi-1 trans st dev	40	7.5	19.0	5.3	-2.2	BETTER
c. Chi-1 gauche plus st dev	74	7.1	17.5	4.9	-2.1	BETTER
d. Chi-1 pooled st dev	158	7.5	18.2	4.8	-2.2	BETTER
e. Chi-2 trans st dev	38	9.4	20.4	5.0	-2.2	BETTER

Figure 3.6: Side-chain parameter analysis for homology modelled malaria DHFR.

superimposed with the structure for human DHFR, and a Cα R.M.S. deviation of 3.6Å was calculated (Figure 3.8).

Active site residues were superimposed (Figure 3.9). Catalytic residues are reasonably similar with the major change being *Glu28Asp* and

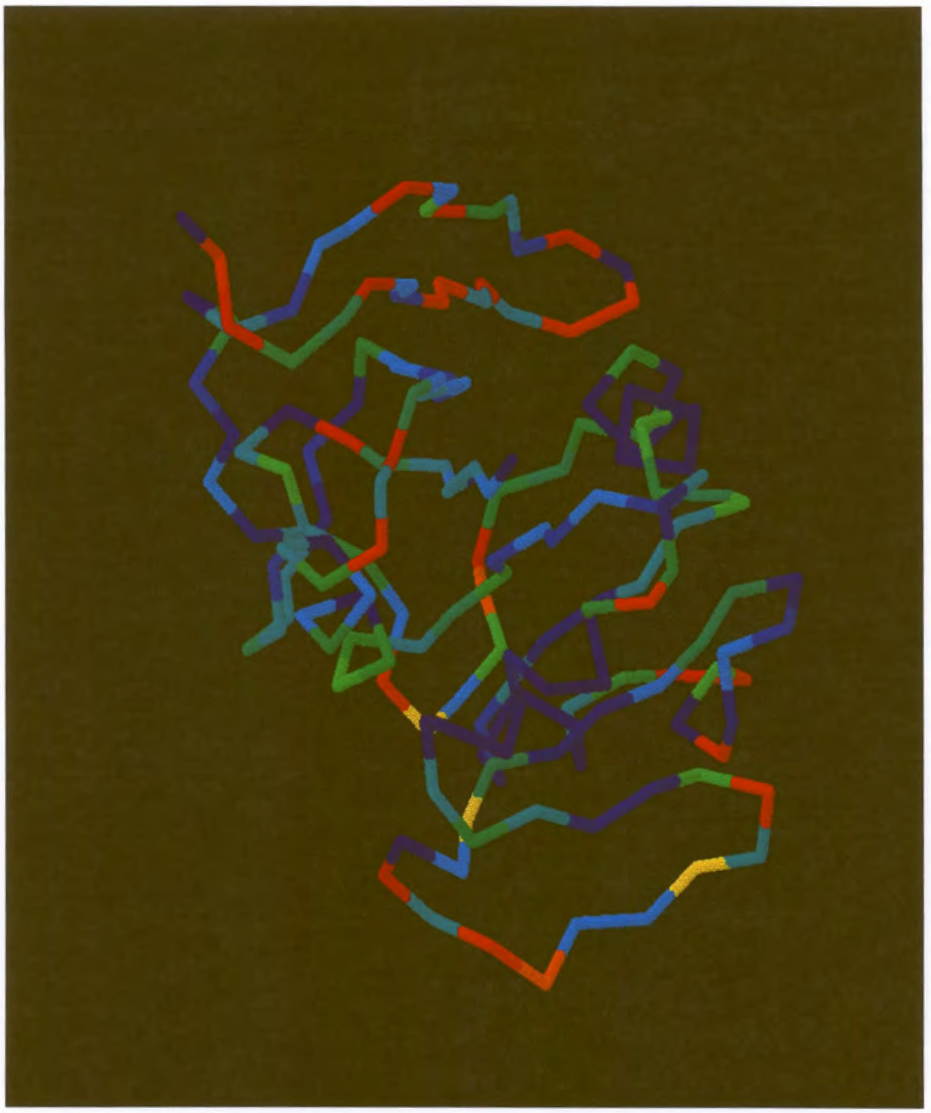


Figure 3.7: Quality score indications of homology-modelled DHFR. Red regions indicate lower quality, green intermediate and blue regions indicate higher quality.

twisted conformations of some side-chains.

The main-chain parameters indicated the correct fold of the malaria DHFR backbone, excepting the putative loop regions excised from the alignments. Side-chain parameters however showed values of inferior quality generating low confidence in the side-chain conformations of many amino acids.

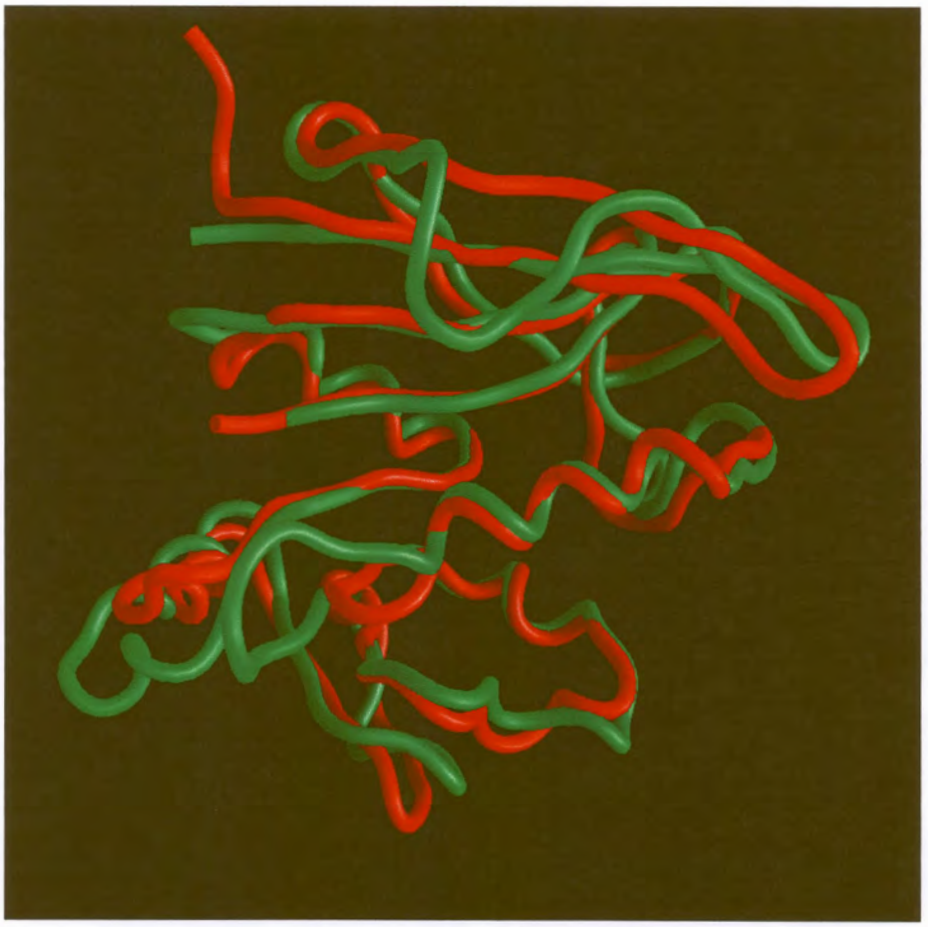


Figure 3.8: Fitted structures of modelled malaria DHFR (red) and human DHFR (green).

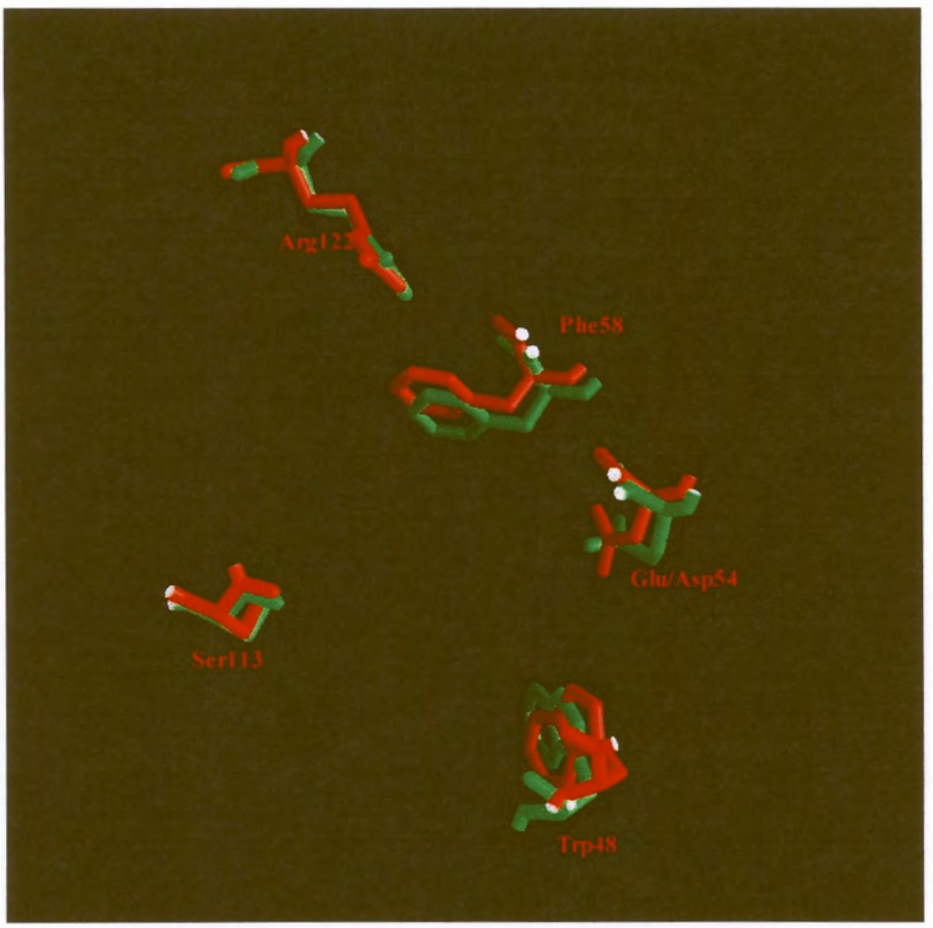


Figure 3.9: Superimposed active site residues of malaria DHFR (red) and human DHFR (green).

3.3.2 Homology modelling of triosephosphate isomerase

The initial amino acid sequence alignment of malaria TIM with those of other species in the Brookhaven Protein Database is shown in Figure 3.10.

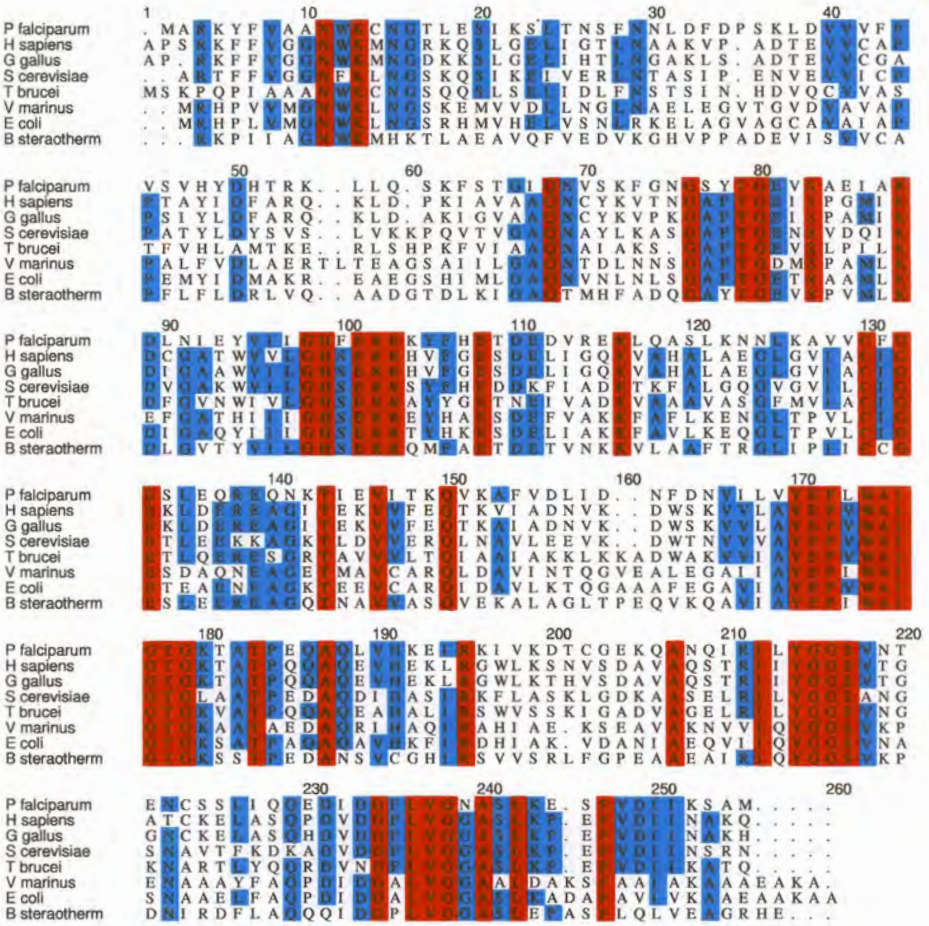


Figure 3.10: Sequence alignments of all TIM structures available in PDB. Homology is indicated in blue and identity in red.

In this alignment, TIM was found to be conserved to a relatively high degree in different species. No large insertions or deletions were observed as was the case with DHFR.

Because of the conserved nature of TIMs, the structure for all TIMs from PDB were used for model building. After model building, the structure was analyzed with the Procheck package (Laskowski *et al.*, 1996). Some results are summarised in Figure 3.11 (results from all

plots not shown).

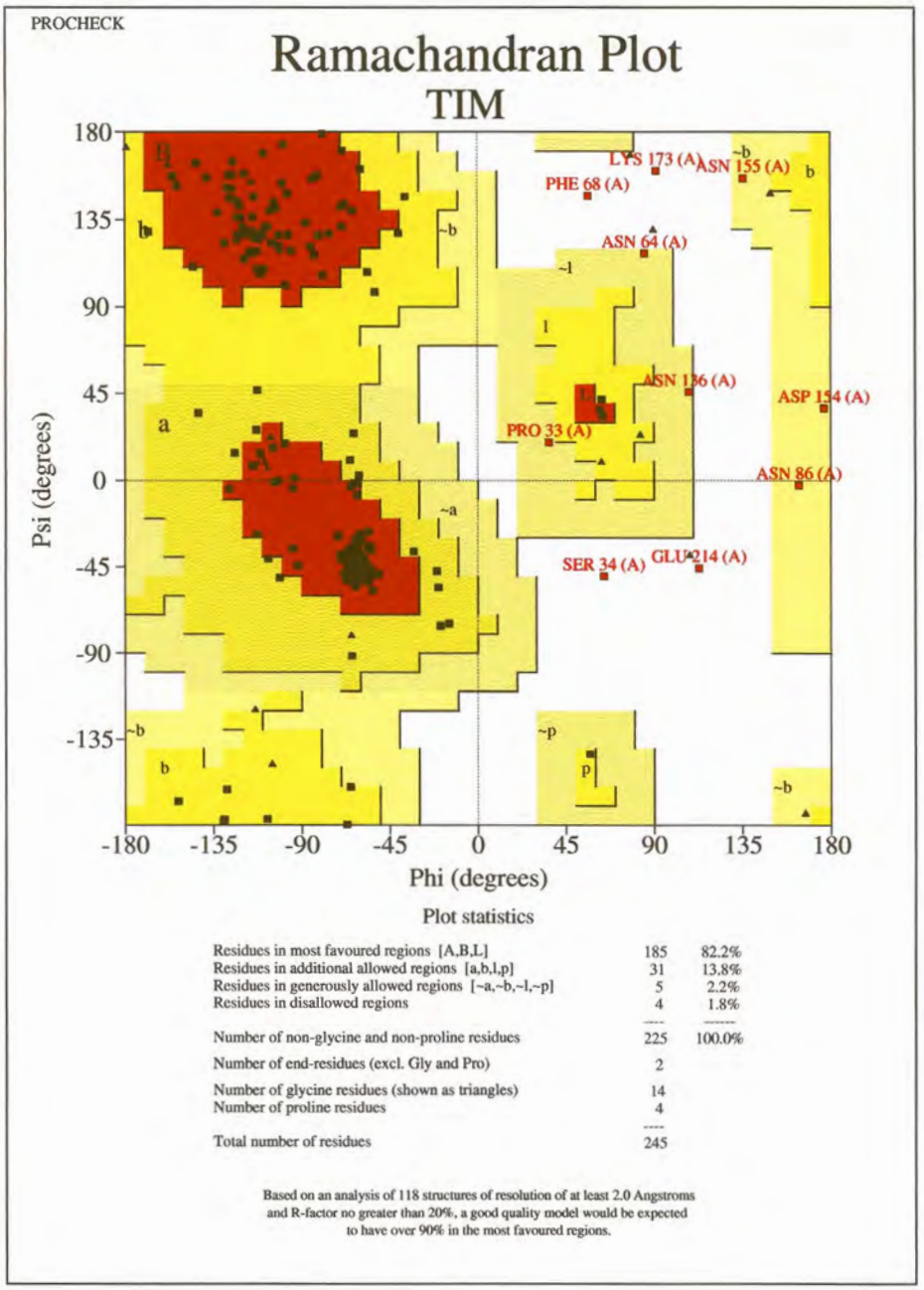
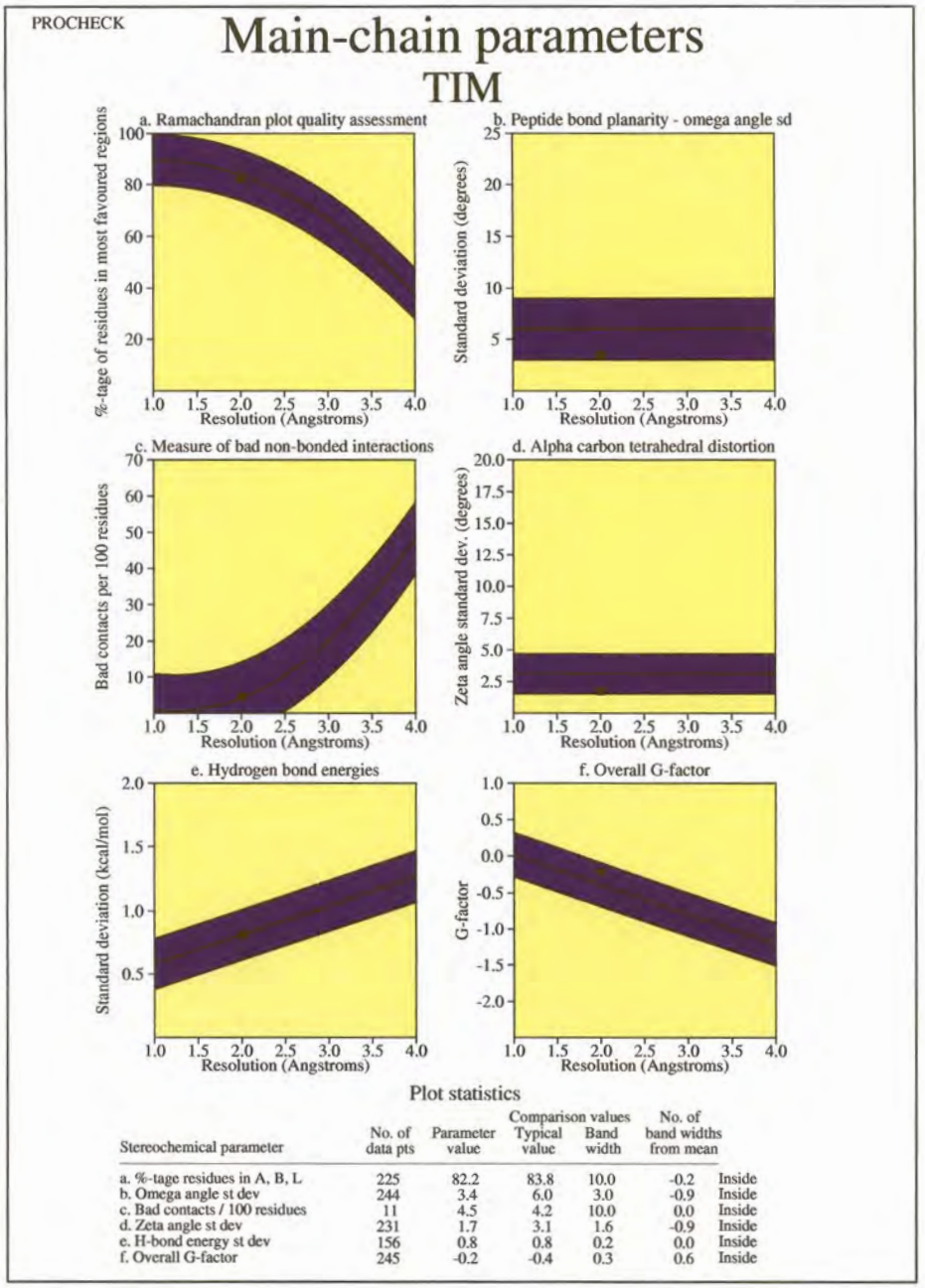


Figure 3.11: Ramachandran plot for the homology model of malaria TIM.

The Ramachandran plot shows 82.2% of residues to be in favoured regions (red), and 13.8% to be in additionally allowed regions (Yellow). Of the remaining residues 2.2% were in generously allowed regions

(beige) and 1.8% in disallowed regions (white). None of the catalytic residues were found in disallowed regions.

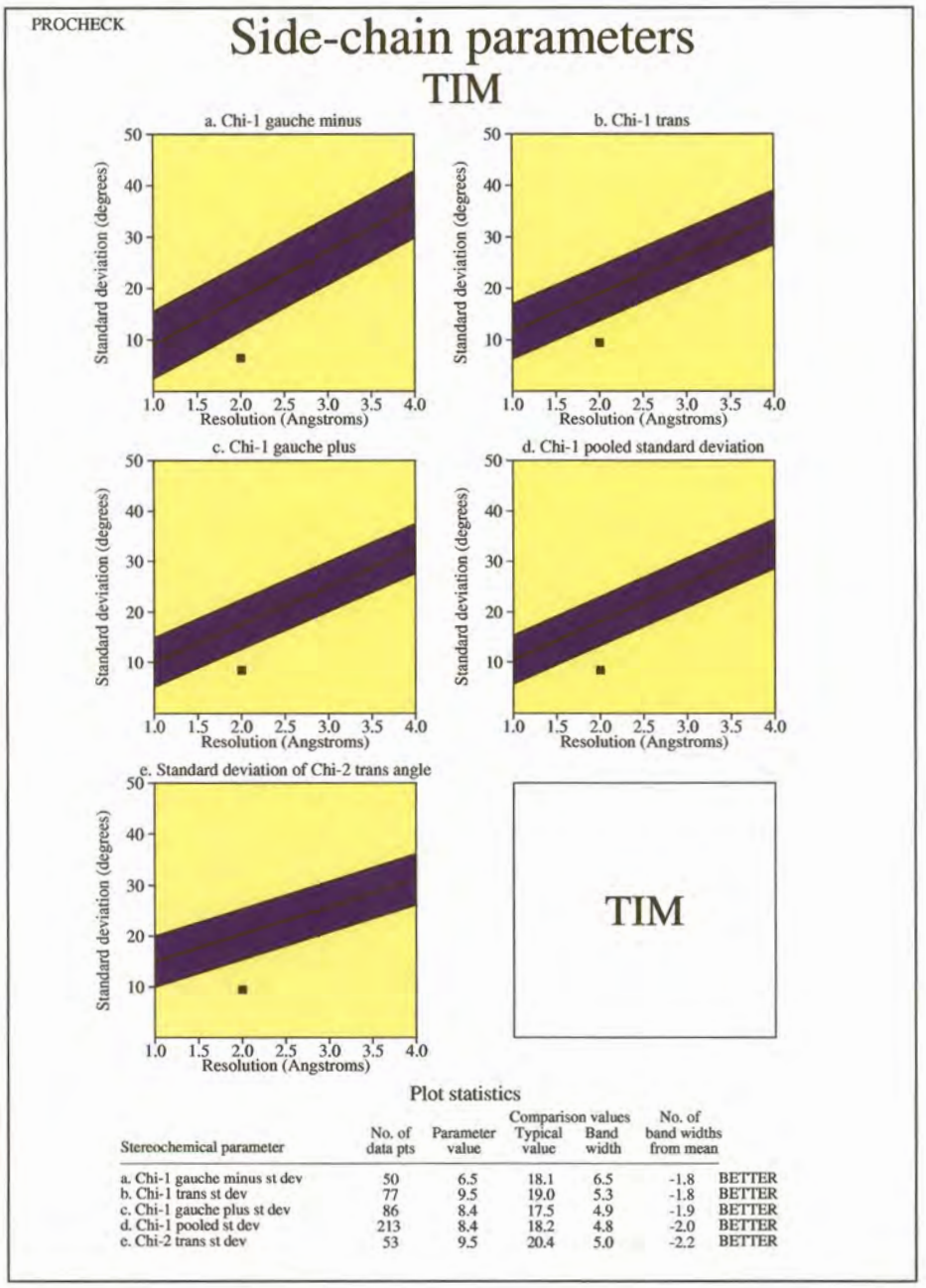


Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	225	82.2	83.8	10.0	-0.2	Inside
b. Omega angle st dev	244	3.4	6.0	3.0	-0.9	Inside
c. Bad contacts / 100 residues	11	4.5	4.2	10.0	0.0	Inside
d. Zeta angle st dev	231	1.7	3.1	1.6	-0.9	Inside
e. H-bond energy st dev	156	0.8	0.8	0.2	0.0	Inside
f. Overall G-factor	245	-0.2	-0.4	0.3	0.6	Inside

Figure 3.12: Main chain parameter analysis for homology modelled malaria TIM.

Further analysis of values from the Ramachandran plot showed the main-chain residues to occur within the allowed deviations from mean

values (Figure 3.12). A bad contact score of 4.5/100 residues was calculated and an overall G-factor of -0.2.



Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. Chi-1 gauche minus st dev	50	6.5	18.1	6.5	-1.8	BETTER
b. Chi-1 trans st dev	77	9.5	19.0	5.3	-1.8	BETTER
c. Chi-1 gauche plus st dev	86	8.4	17.5	4.9	-1.9	BETTER
d. Chi-1 pooled st dev	213	8.4	18.2	4.8	-2.0	BETTER
e. Chi-2 trans st dev	53	9.5	20.4	5.0	-2.2	BETTER

Figure 3.13: Side-chain parameter analysis for homology modelled malaria TIM.

Side-chain parameters were not of high quality but were generally closer to the accepted deviation from mean than those of DHFR (Figure 3.13).

A three-dimensional diagram was prepared indicating the quality values for the model by different colours (Figure 3.14).

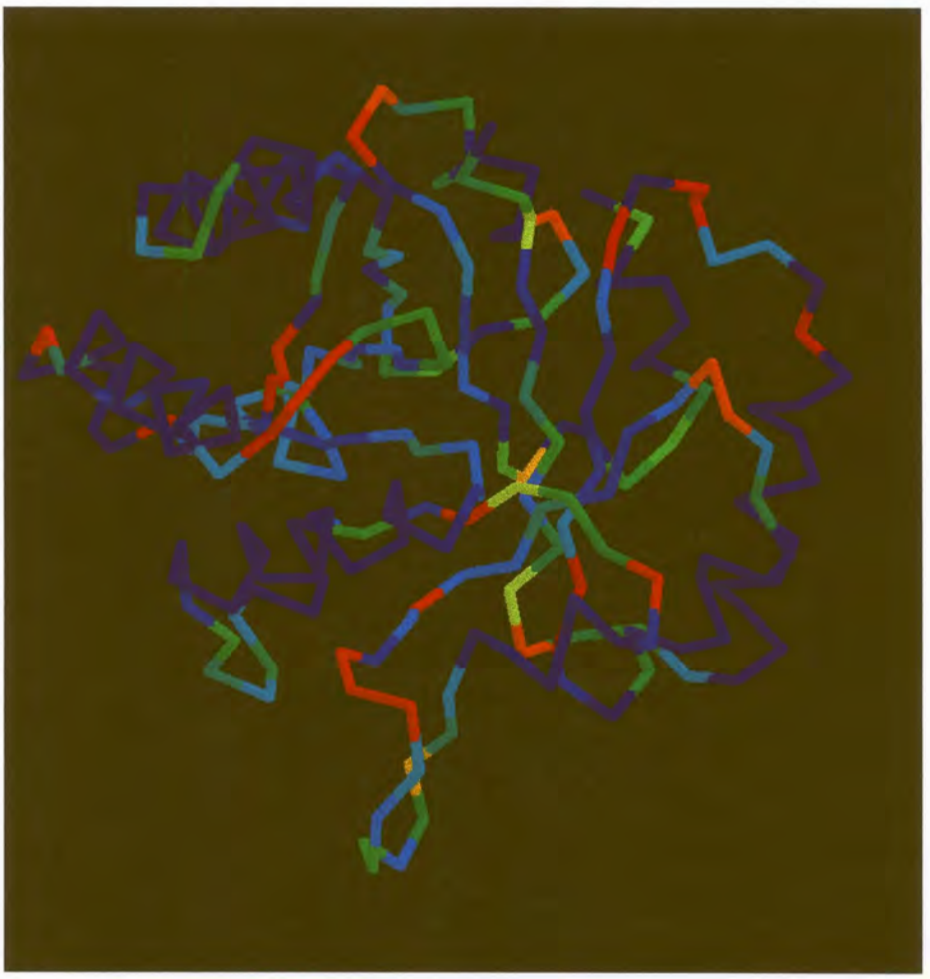


Figure 3.14: Quality score indications of homology-modelled TIM. Red regions indicate lower quality, green intermediate and blue regions indicate higher quality.

Much less red-colored regions are visible than in the case of DHFR. Low quality scores were mostly found in some loops at the surface of TIM. The homology structure was superimposed with the structure for human TIM to investigate the possibility of preparing selective inhibitors, and a R.M.S. deviation of 0.893Å was calculated (Figure 3.15).

Small deviations were visible in some helices and in loops at the surface of the TIM molecule. Active site residues were superimposed (Figure 3.16). Catalytic residues were nearly identical except for the *Ser95Phe* mutation and some slight rotation of the *His94* side-chain.

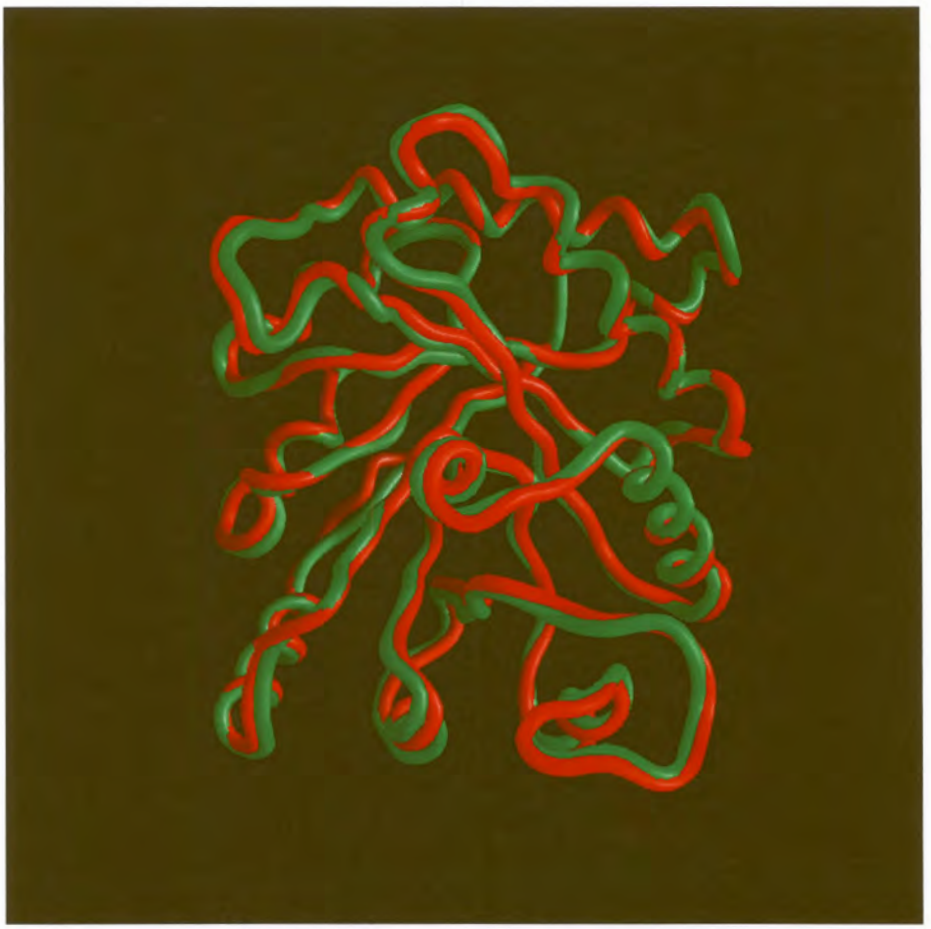


Figure 3.15: Fitted structures of modelled malaria TIM (red) and human TIM (green).

The main-chain parameters are of high quality, possibly due to the high number of templates used for model building and to the very high degree of conservation of the TIM family fold. The side-chain parameters are of lower quality, correlating with the observed lower degree of amino acid identity and high degree of fold homology found for the different TIMs.

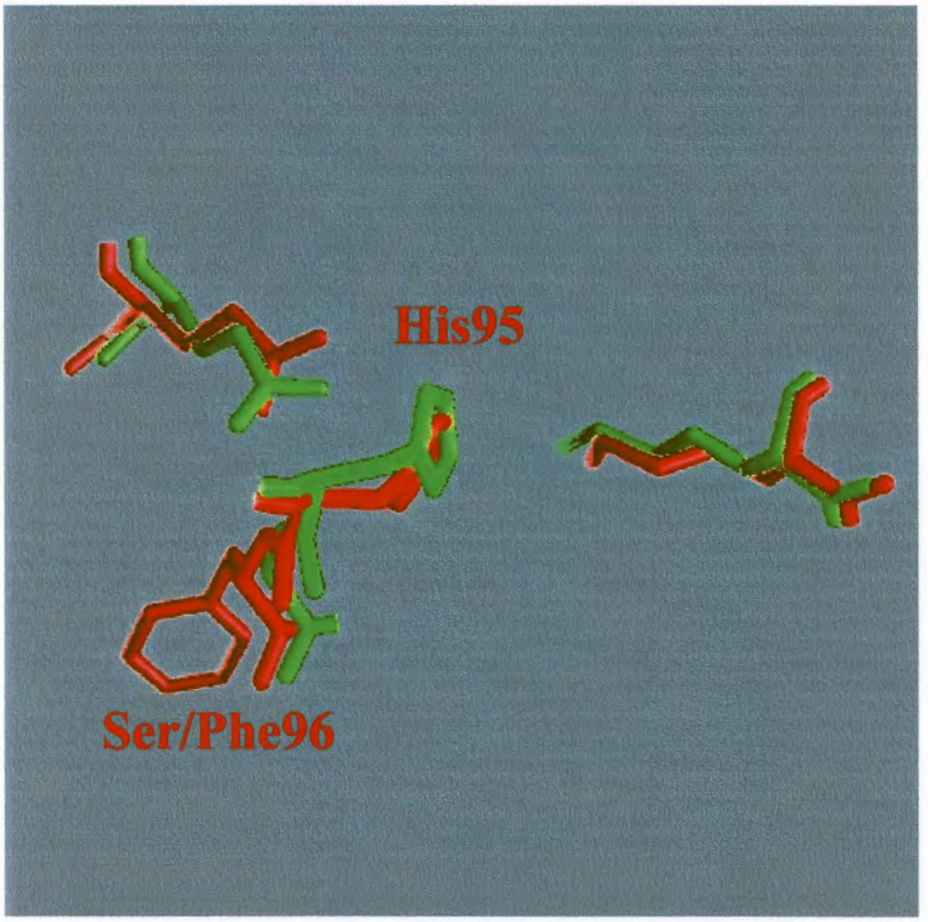


Figure 3.16: Superimposed active sites of malaria TIM (red) and human TIM (green).

3.4 Discussion

The homology modelling of *P. falciparum* DHFR was complicated by the presence of a 27 residue insertion near the N-terminal region of DHFR, as well as a 23 residue insertion near the C-terminal region. Structures from *E. coli* and *L. casei* were chosen as template, and the sequences were re-aligned. The two insertions did not seem to be involved in active site regions, and insertion regions were excised from the malaria sequence in order to facilitate model building for practical reasons, although this is not an ideal situation since the functions of these regions have not been ascertained. A recent publication by Lemcke *et al.* (1999) also addressed the preparation of homology models for malaria DHFR, but no structural models were made available. They suggested the inserted regions occur as loops pointing away from the surface of the proteins, but could not conclusively offer proof regarding the loop conformations, and found relatively low quality values for these regions. These loops are located well away from the active site cleft, and were proposed not to affect the model in terms of active site investigation. They used the conserved regions as template for model building, and added the loops afterwards from libraries of preferred side-chain conformations. The N-terminal extension has been proposed to improve contact between the two domains of DHFR. Lemcke *et al.* (1999) used their model to propose a mechanism for the influence of the Ser108Asn point mutation on pyrimethamine resistance, but could not yet explain the mechanism of the Asn51Ile mutation. Toyoda *et al.* (1997) prepared a homology-based model of DHFR by methods not yet published, and this was used for ligand design purposes (See Chapter 4). They were able to prepare inhibitors with micro- to nanomolar affinities in this way.

An alternative to modifying known drugs would be to find compounds with structures differing from the known inhibitors such as pyrimethamine, cycloguanil and methotrexate, thus decreasing the chance of the rapid development of drug resistance at the same sites. The validity of constructing a homology model of DHFR without the adjoining junction and TS domains remains debatable. *In vitro* studies have shown that DHFR will function normally without these domains (Prapunwattana *et al.*, 1996)(Sano *et al.*, 1994). However, the possibility that excision of insertion loops may affect the overall fold of the protein and also the regions facing the active site, can not be excluded.

The homology modelling of malaria TIM was less difficult than DHFR, as the sequences for TIM are very similar in all species. Quality reports of homology modelled TIM indicated satisfactory main-chain and acceptable side-chain values. Although the sequence identity is not very high, homology is up to 98% (chicken) and no large insertions or deletions are present. Near the end of this study the X-ray structure for *P. falciparum* TIM was published (Velanker *et al.*, 1997). Coordinates were obtained from the authors, and the homology and X-ray structures were compared. Cα backbone comparisons of homology modelled TIM with the X-ray structure showed a Cα R.M.S. deviation of 1.5Å (Figure 3.17). Some differences were visible in coil and loop regions.

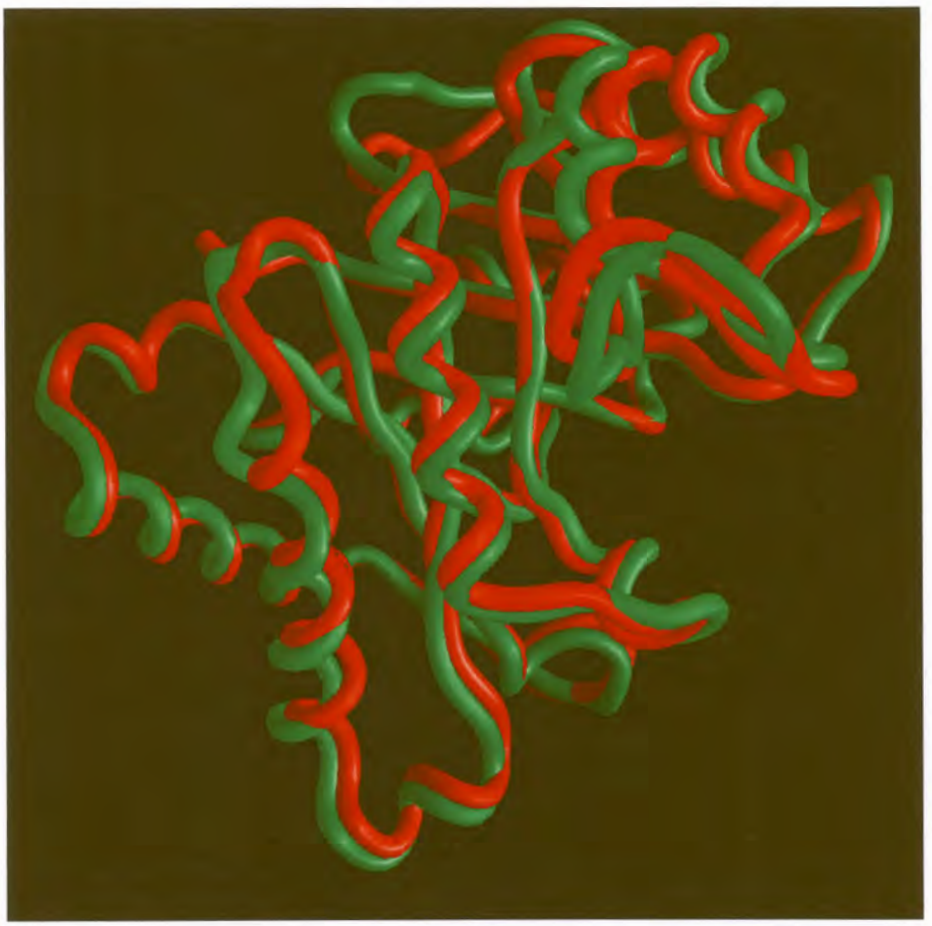


Figure 3.17: Superimposed Cα backbones of homology modelled TIM (red) and the X-ray structure of TIM (green). A R.M.S. deviation for carbon-α of 1.5Å was found.

A detailed analysis of the active sites showed the model to be nearly

identical to the X-ray structure, except for the rotation of Phe96 (Figure 3.18). This residue did however attain a most favoured rating in the model according to the Ramachandran plot (not shown).

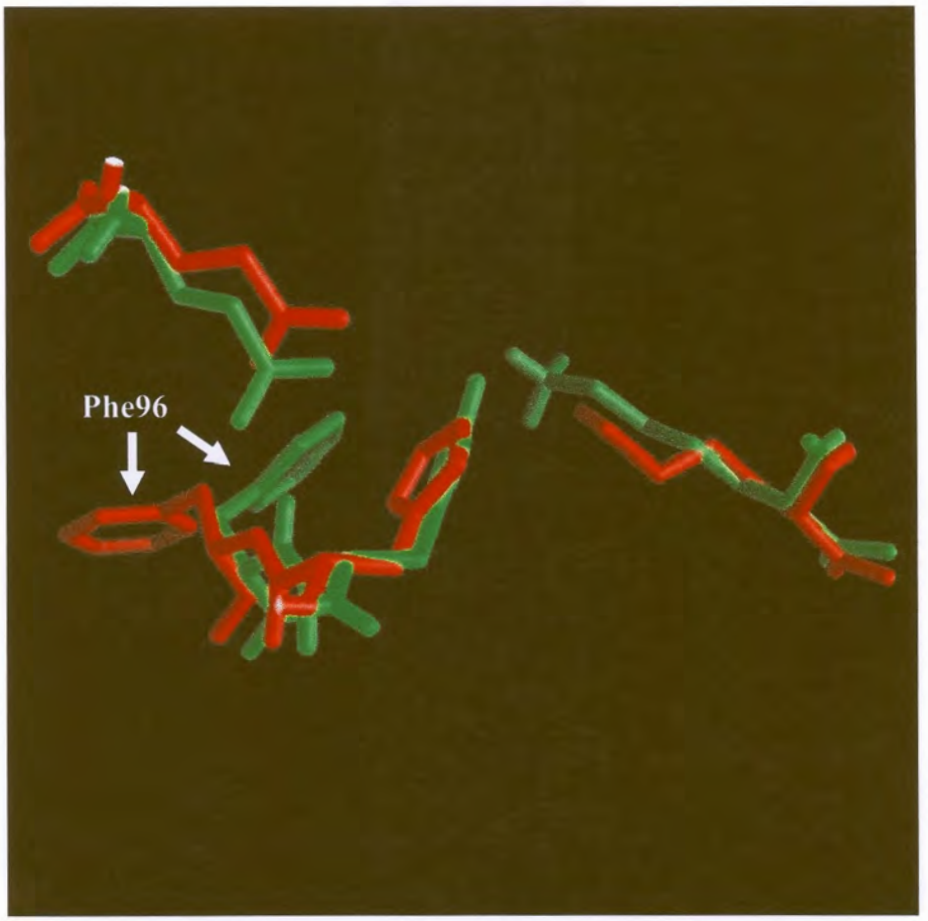


Figure 3.18: Superimposed active site residues of homology modelled TIM (red) and the X-ray structure of TIM (green). The only major difference was the rotation angle of Phe96.

Although the homology model of malaria TIM would probably have been of high enough quality for ligand design, the X-ray structure (1YDV) was used in further docking studies.

Chapter 4

Ligand discovery for malaria triosephosphate isomerase

4.1 Introduction

With current increases in world-wide malaria resistance, investigations into new inhibitors as lead drug compounds are needed. The design of new enzyme inhibitors is a challenging task. With the onset of computerised modelling in drug discovery the development times of successful drugs have decreased dramatically. All proteins bind their ligands in specific conformations. The protein and ligand combining sites determine both the affinity and the specificity of the interaction. In this interaction various forces need to be taken into account. Hydrogen bonds between peptide bonds are perhaps the greatest force in proteins, organising the proteins into secondary structure elements such as α helices and β sheets. Apolar and hydrophobic interactions also contribute to protein stability, together with side-chain to backbone hydrogen bonding.

A series of docking packages have been designed. Some are meant to determine *de novo* structures fitting into active sites, while others are designed for the scoring of existing molecules. Two main types of approaches will now be described.

The AutoDock package is optimised for the automated docking of flexi-

ble small ligands to receptors. AutoDock uses a Monte Carlo simulated annealing approach for the configurational exploration with fast energy calculation based on molecular affinity potentials (Goodsell *et al.*, 1996). Atomic affinity potentials are precalculated as described by Reynolds *et al.* (Reynolds *et al.*, 1989). The protein is embedded in a 3D grid and a probe atom is placed at each grid point. The energy of this single atom with the protein is then assigned to this grid point. This is performed for each type of atom in the ligand molecule, as well as a grid of electrostatic potentials. The docking simulation is carried out by performing a random movement of the ligand through the static protein. At each step a small random displacement is applied to each of the degrees of freedom of the ligand. If the energy is lower than that of the previous configuration, it is accepted. If higher, it is rejected based on a defined limit or a probability expression. Simulated annealing allows exploration of a complex configurational space. A range of molecular interactions may be taken into account such as dielectrics, potential functions, etc. This approach has been used for the docking of ligands to fXa inhibitor (Rao and Olson, 1999), the cannabinoid receptor (Mahmoudian, 1997), antibodies (Gamper *et al.*, 1996) and many other proteins.

The DOCK package (Gschwend *et al.*, 1996) used in this study is designed to find favourable orientations of a ligand in a receptor. The ligand orientation in the binding site is broken down into a series of steps. First a potential site of interest in the protein is identified. Within this site, points are identified where ligand atoms may be located. This is done by generating a set of overlapping spheres to fill the site. This limits the potential number of orientations in the active site. Spheres may overlap each other but not the protein surface. To orient a ligand within the active site some of the sphere centers are matched with ligand atoms, generating many such sets. A set of atom-sphere pairs are used to calculate the orientation of the ligand at the site of interest. Various filters are used to eliminate conformations which would generate poorly scoring orientations, thus only a small subset of possible ligand orientations are generated and scored. The spheres do not take chemical matching into account. These properties may then be introduced by matching characteristics such as hydrogen-bond donors, hydrophobes, electro-positiveness, electro-negativeness, etc. In this way the number of unfavourable operations which are generated and scored may be further reduced. A critical point set may be defined in-

corporating known interactions. This method has been used with great success in cases such as *P. carinii* DHFR (Gschwend *et al.*, 1997), influenza virus hemagglutinin (Hoffman *et al.*, 1997) and double-strand RNA systems (Chen *et al.*, 1997).

Before scoring, a bump filter is applied which eliminates orientations in which ligand atoms occupy space already in use by the protein. The orientation of the ligand is then evaluated with a shape scoring function and/or a function approximating the protein-ligand binding energy. Evaluations are performed on scoring grids to minimise computing time. This approach stores the protein contributions to the score only once according to the grid, and retrieves them as necessary. The shape scoring function is an empirical function similar to the van der Waals attractive energy. The shape score is generated by summing the receptor terms from the grid point nearest to each non-hydrogen ligand, thus the shape score is determined by the position of each ligand atom on the shape scoring grid. The binding energy of the ligand-protein interaction is approximated by the sum of the van der Waal attractive, van der Waal dispersive and Coulombic electrostatic energies. To generate the energy score, ligand atom terms are combined with receptor terms from the nearest or virtual interpolated grid point. Thus the energy score is determined by both the ligand atom types and ligand atom positions on the grid.

The 3D Available Chemicals Database is a good source of ligand structures but is expensively priced. The Cambridge Crystallogaphic Database contains accurate structures but many of them play no role in biological systems. It was decided to screen against the National Cancer Institute (NCI) 3D database. The advantages are that this database contains compounds used in cancer studies, thus the biological side effects of many of these molecules are known. The database is also freely available without any cost. However, the 3D database was theoretically generated with the Corina package (Gasteiger *et al.*, 1990), thus altough the structures are energetically optimal, they are not always possible to synthesise in these conformations. Also, many of the compounds have been specifically synthesized and are not commercially available.

As DHFR was not expressed in an optimised form, and the DHFR homology models were of lower quality, DHFR was not employed for ligand screening. Also, the newly available X-ray structure of TIM increased the chances of finding successful inhibitors.

The following aims were set for this study:

- The screening of a 3D molecular database against the active site region of malaria TIM
- The *in vitro* screening of putatitive inhibitors against recombinant TIM
- The screening of putatitive inhibitors against malaria cultures

4.2 Methods

The X-ray structure (1YDV) for malaria TIM (Velanker *et al.*, 1997) was used for all docking studies. Of the dimer A- and B-chains, the A chain was used, and cleaned for waters and other ions. SYBYL (Tripos Associates, St. Louis) was used for correction of the N- and C-terminal residues and for hydrogen addition and charge assignment. The structure was subsequently converted to MOL2 format. A grid box was manually added to define the region of interest, and the GRID module was used to define the steric and electrostatic properties at each grid point to enhance rapid scoring of ligand orientations during the DOCK run. The molecular surface was created with the QCPE MS package (University of California, Berkeley). The SPHGEN module was used for the calculation of spheres filling the site, and the sphere site points were optimised by visual inspection. The NCI-3D database (Milne *et al.*, 1994) was converted from SDF to MOL2 format, and SYBYL was used to assign atom types and charges. The docking run was performed with the DOCK 4 (Gschwend *et al.*, 1996) package on a Silicon Graphics Challenge S system.

Inhibition kinetics of ligands to recombinant malaria TIM was assayed as described in paragraph 2.3.2. Inhibitors were added in the 10 to 250μM range and assayed for changes in TIM activity. Results were used to prepare inverse reciprocal plots of reaction rate against substrate concentration.

To test the effect of the inhibitors on parasites, malaria cultures were incubated with inhibitors at concentrations of 1000μM, 100μM and 10μM for a period of 24 hours. Cultures were subsequently stained with thiazole orange according to Schulze *et al.* (1997) and parasitemia measured by flow cytometry (Coulter FACSScan, Beckman-Coulter, California).

4.3 Results

After conversion of the database from SDF to MOL2 format and assignment of charges and hydrogens in SYBYL, 107,955 compounds were tested for docking. After removing compounds with high bump scores (thus not being able to fit because of their size), 105029 compounds were suitable for docking in the defined region. The region of interest was optimised by visual inspection to surround the active site region of malaria TIM. The superimposed structure of human TIM bound to PGA was used to determine the center of the TIM active site region (Figure 4.1) and to select a bounding box.

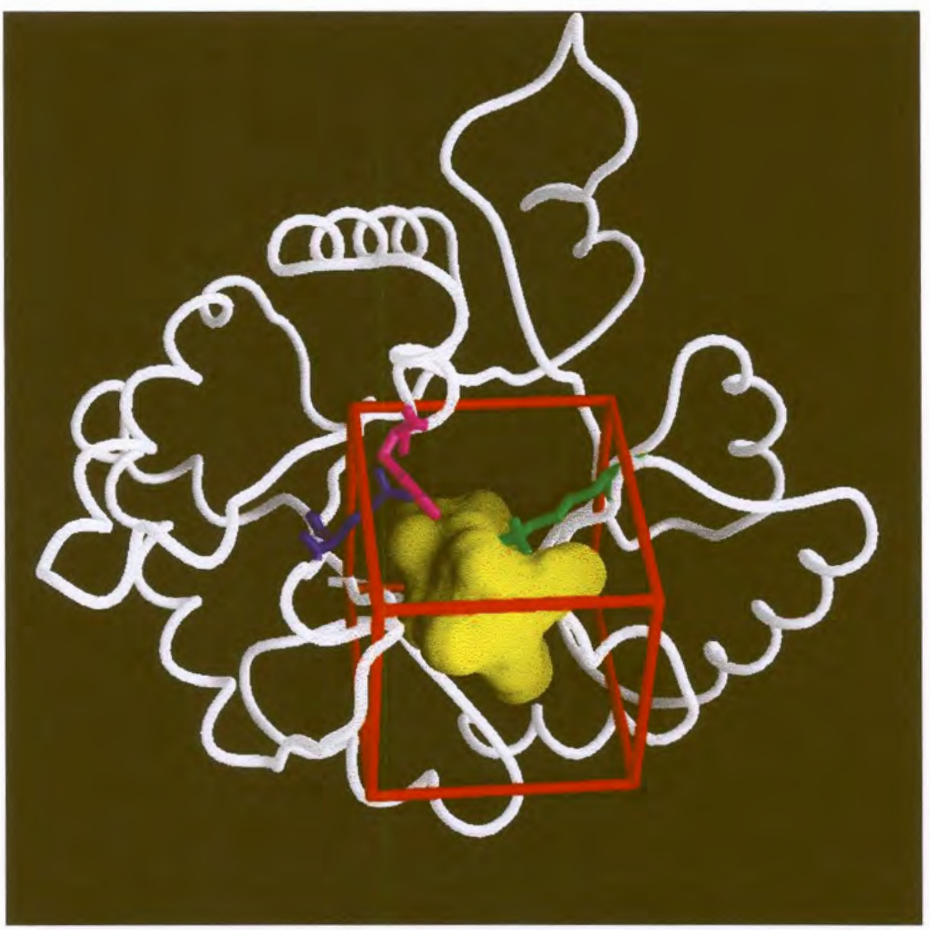


Figure 4.1: Region of malaria TIM surrounding the active site used for ligand docking. The active site was determined by the binding position of the inhibitor PGA. Yellow spheres indicate the active site surface, and the region of interest is bound by a red box. Active site residue *Lys12* is indicated in green, *His96* in magenta and *Glu165* in blue.

The spheres for the docking space were optimised by visual inspection before use (Figure 4.2) and spheres leading away from the desired region were removed.

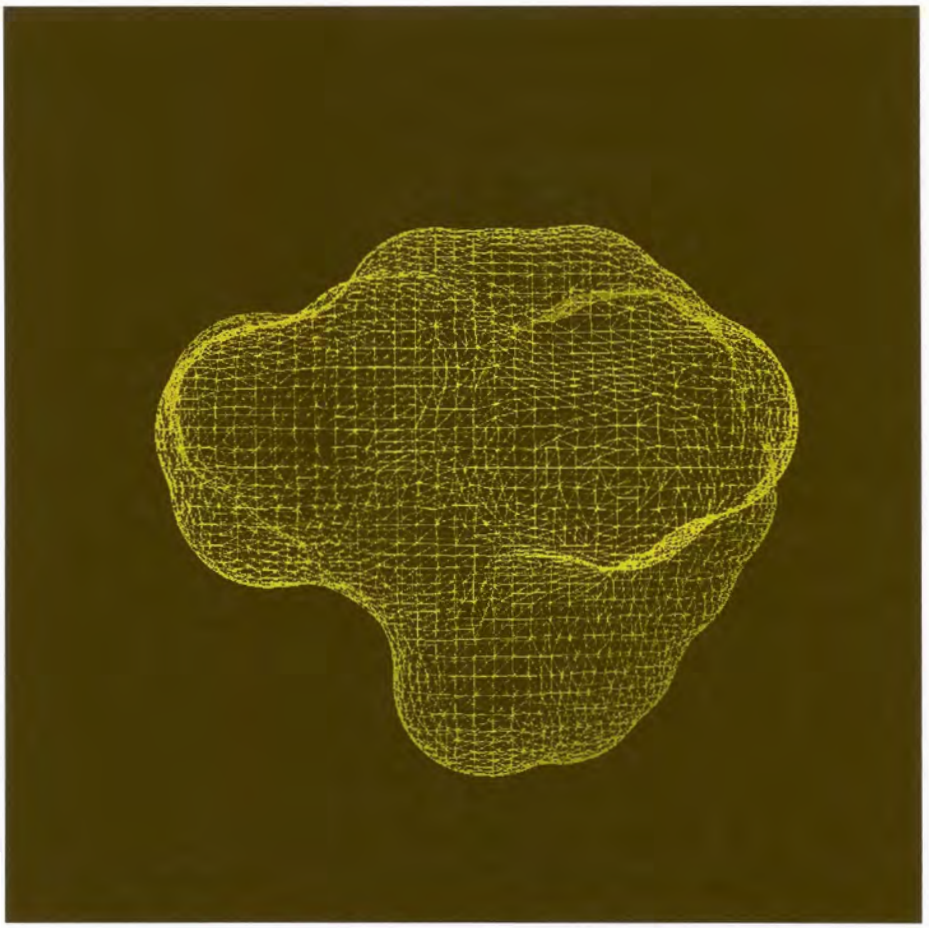


Figure 4.2: Detailed view of spheres defining the inverse of the active site cavity.

Computer screening took approximately 500 hours. The top 100 hits were saved, and manually inspected for suitability in terms of size, physical properties, solubility, etc. Various permutations of the same molecules were discarded, for example compounds differing only in the bound salt ion, etc. Hits were further refined according to commercial availability.

The following compounds were chosen for inhibition studies (Table 4.1 and 4.2):

The commercially unavailable compounds are being synthesized by

Table 4.1: Commercially available compounds tested for TIM inhibition.

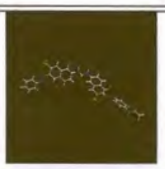
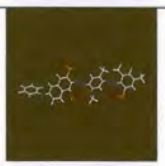
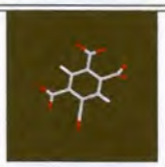
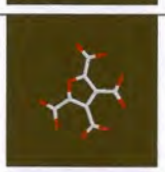
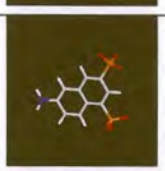
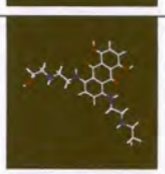
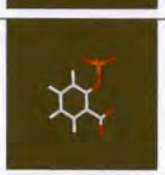
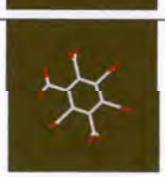
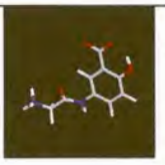
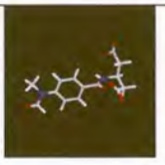
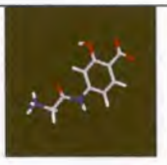
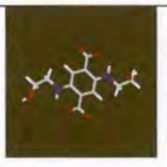
Compound	Structure	Commercial availability	Score
Direct Red 23		Yes	-38.31
Direct Violet 51		Yes	-36.89
Pyromellitic acid		Yes	-36.05
Furanetetracarboxylic acid		Yes	-35.59
Amino I acid		Yes	-35.47
Mitoxanthrone		Yes	-35.11
Tricarballylic acid		Yes	-34.66
Salicyl phosphate		Yes	-34.49
Mellitic acid		Yes	-33.72

Table 4.2: Synthesized compounds tested for TIM inhibition.

Compound	Structure	Commercial availability	Score
Bicyclo[2.2.1]heptane-2,3,5,6-tetracarboxylic acid		No	-41.31
5-(2-aminoacetamido)-salicylic acid		No	-37.37
3-[[4-(formylmethylamino) benzoyl] amino]-pentanedioic acid		No	-36.93
4-(2-aminoacetamido)-salicylic acid		No	-36.66
2,5-bis[(2-hydroxyethyl)amino]-terephthalic acid		No	-36.34

Prof. T van Ree at the University of Venda but were not yet available at the time of writing this thesis.

Inhibition studies on recombinant purified malaria TIM were performed with DHAP as substrate. Of the commercially available inhibitors, only Direct Red 23 and Direct Violet 51 showed any inhibition of TIM activity at concentrations lower than 100μM. Kinetic studies of these two compounds were performed to determine the mode of inhibition (Figure 4.3).

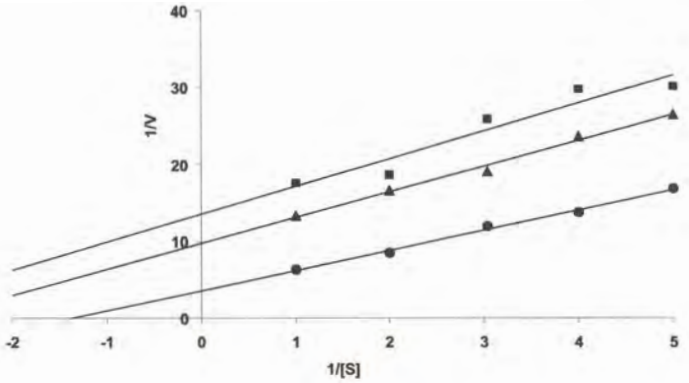


Figure 4.3: Inhibitor kinetic plot for Direct Red and malaria TIM. Squares indicate a concentration of 60μM, triangles 50μM and circles 40μM.

For Direct Red, a mixed type inhibition was found. Whereas competitive inhibition would yield lines converging to the same 1/[V], and uncompetitive inhibition would yield parallel lines, the plots indicated an inhibition mode similar to uncompetitive inhibition, with slightly varying angles to the parallel lines. Direct Violet showed similar values (results not shown).

Inhibition of malaria growth in culture was measured by thiazole orange staining and flow cytometry (Figure 4.4). This method discerns between the malaria parasites and red blood cells based on the fluorescence of genomic DNA when stained with thiazole orange.

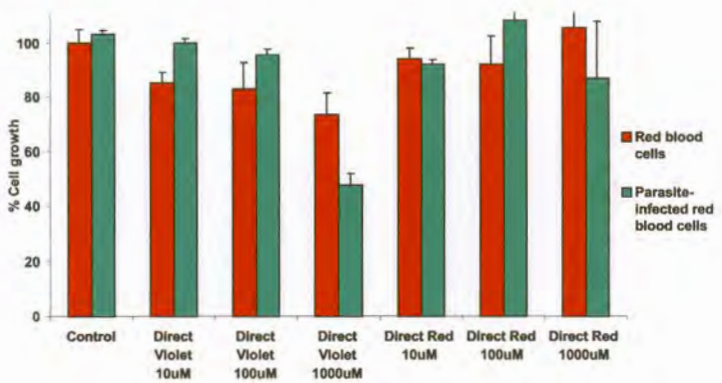


Figure 4.4: Parasite growth in the presence of Direct Red and Direct Violet. Red bars indicate red blood cell growth and cyan bars indicate parasite growth.

Direct Violet showed a reduction in parasitemia at concentrations of 1000μM, but not at lower concentrations. However, a similar decrease was found for red blood cells. No specific effect on malaria parasites could thus be shown. Direct Violet did not yield any significant differences. Other questions remain to be answered, such as the penetrative ability of these compounds through the red blood cell membrane, and subsequently through the parasite membrane.

Investigation into the binding modes of Direct Red and Direct Violet showed the compounds to be bound at the active site of TIM, with one end buried and the other protruding into the solvent (Figure 4.5).

Analysis of intermolecular contacts by LIGPLOT showed common contacts involved in the binding of both compounds (Figure 4.6, 4.7).

Both molecules bound with a diphenyl-phenyl type conformation in the protein, but not at the same depth. A common substructure bound with the exact same conformation is visible when Direct Red and Violet are superimposed (Figure 4.8).

Of interest are the common contacts that both compounds have with actve site residues *Phe96*, *Glu165* and *Val212* . Especially *Phe96* and *Glu165* play a role in substrate contact in TIM from many different species. Lead compounds may possibly be synthesized based on the common structure of Direct Red and Violet showing contact with the above residues.

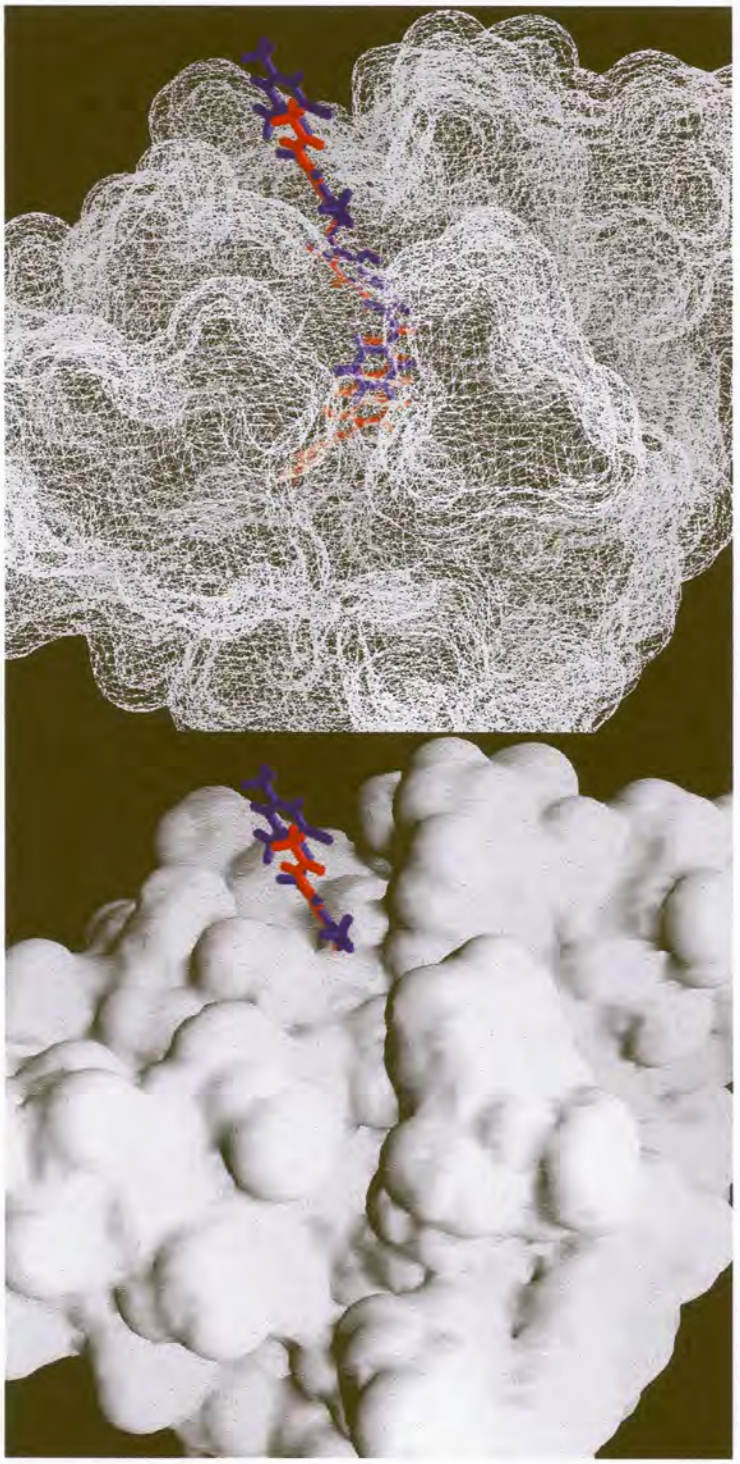
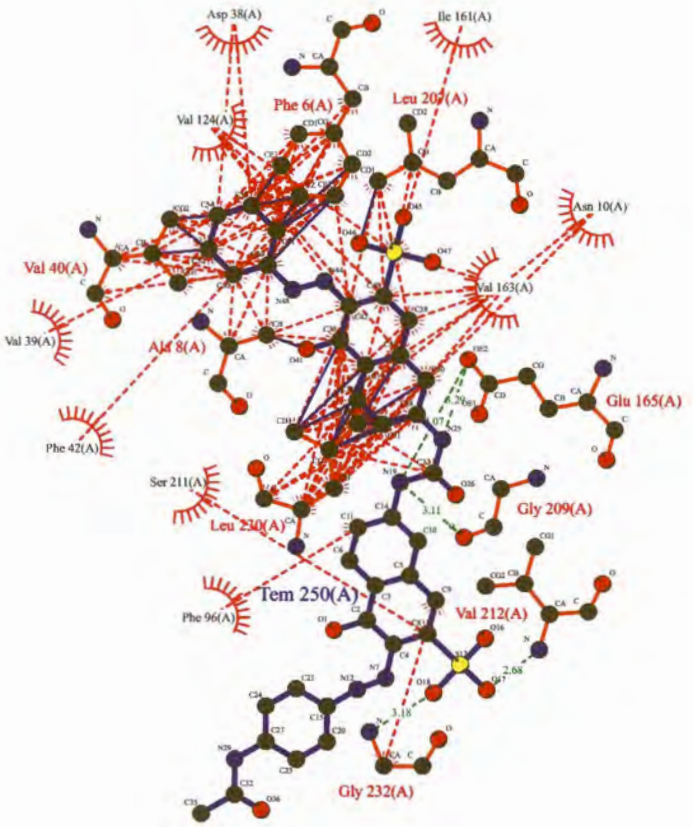


Figure 4.5: Direct Red and Violet bound in the active site of *P. falciparum* TIM. An accessibility surface is shown as mesh (top) or as solid (bottom).



Key

Ligand bond

Non-ligand bond

Hydrogen bond and its length

His 53 Non-ligand residues involved in hydrophobic contact(s)

Corresponding atoms involved in hydrophobic contact(s)

Direct Red

Figure 4.6: Detailed contact map of the malaria TIM-Direct Red complex.

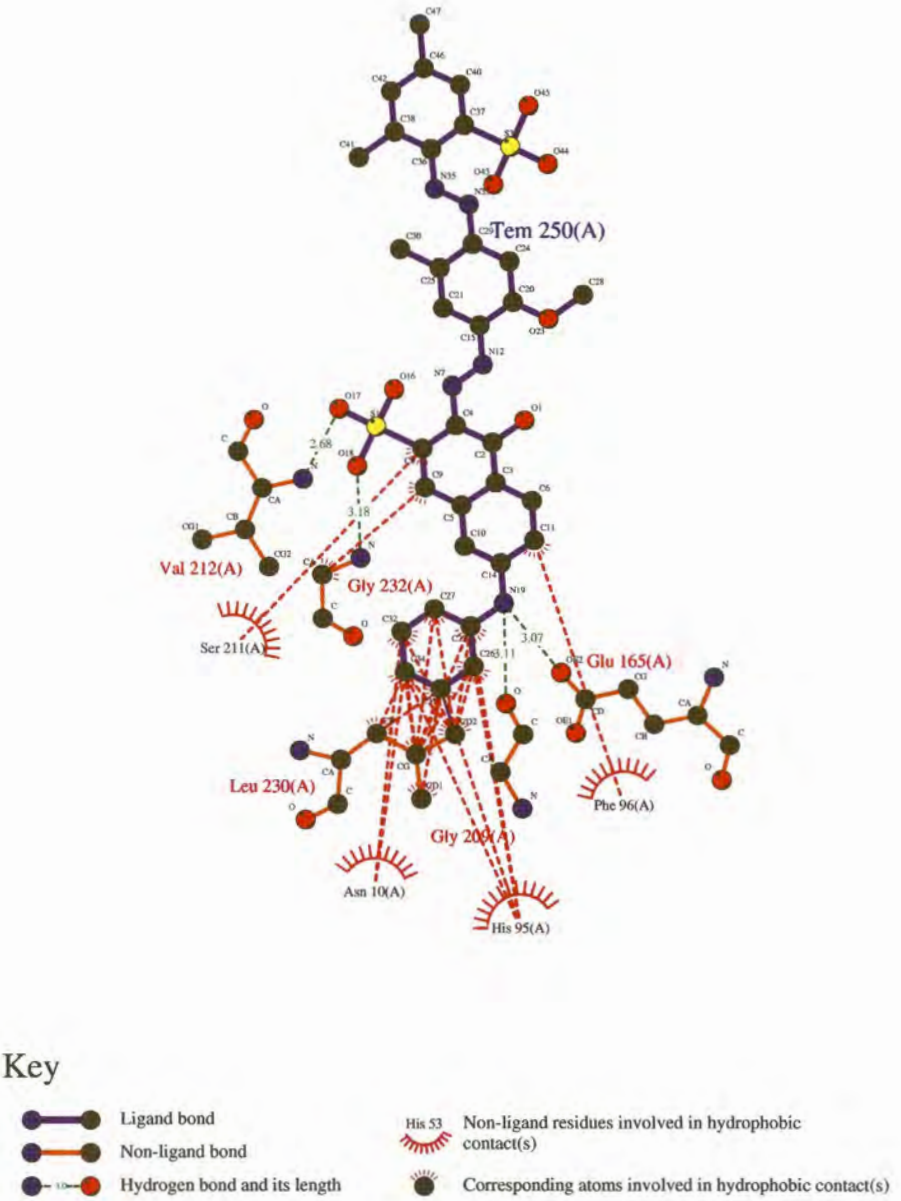


Figure 4.7: Detailed contact map of the malaria TIM-Direct Violet complex.

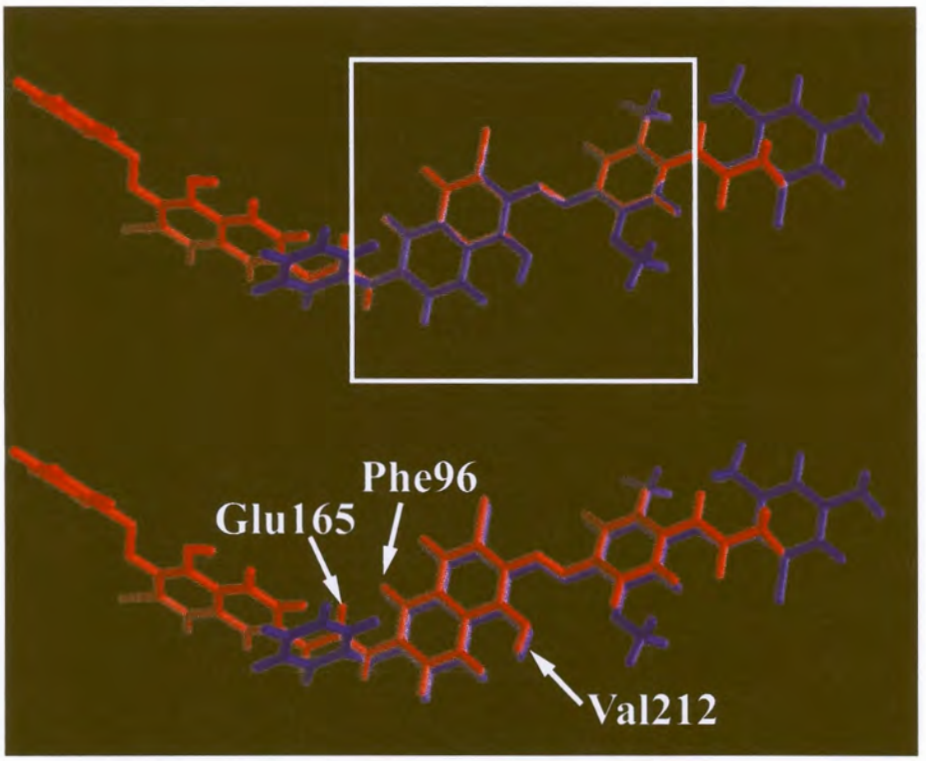


Figure 4.8: Comparison of the spatial positions of bound Direct Red and Violet. The common structural area is indicated by a white box. *Phe96*, *Glu165* and *Val212* show common contact points on both molecules. The top figure shows the exact orientation, while the bottom figure has been shifted slightly to distinguish the two molecules.

4.4 Discussion

Triosephosphate isomerase has long been a target for a series of inhibitors. Crystallisation studies were initially performed with bound PGH which is a substrate analogue (Zhang *et al.*, 1994). Verlinde *et al.* (1992) prepared crystals of *T. brucei brucei* TIM with the competitive inhibitor N-hydroxy-4-phosphono-butanamide, finding that the active site flexible loop was locked in a completely open conformation. TIM has been extensively studied in the parasite *T. brucei brucei* for the design of putative inhibitors. Hydrophobic cyclic hexapeptides were synthesized, and were found to be active and selective inhibitors of TIM (Kuntz *et al.*, 1992). In a study by Schnackerz *et al.* (1991), 3-chloroacetol phosphate (CAP) was shown to be bound to the active site by reaction with *Glu165* to irreversibly fill the binding site of one subunit. The reversibly binding transition state analogue PGA, was used to probe the remaining vacant active site of the heterodimer and showed independent binding at the two active sites. Lolis *et al.* (1990) showed that the inhibitor formed hydrogen bonds to the side-chains of *His95* and *Glu165*. The latter hydrogen bond confirmed that *Glu165* was protonated upon PGA binding. Conformational changes were observed: the side-chain of *Glu165* moved over 2Å and a 10-residue flexible loop moved over 7Å to close over the active site like a lid. The ferrate anion, an analog of orthophosphate anion, was also shown to rapidly inactivate triose phosphate isomerase from chicken muscle. The inactivation can be prevented by the presence of competitive inhibitors (Steczko *et al.*, 1983).

Velanker *et al.* (1997) described the active site of malaria TIM, and found the structures of malaria and human TIM to be very similar. This makes the design of species-specific inhibitors highy challenging. The high specificity of TIM makes this task even more difficult. The only significant differences in the active site are at *Cys13* and *Phe96*. Also, a charged surface at *Glu183* is replaced with the hydrophobic *Leu183*, which has been proposed to play a role in membrane attachment. The other potential site for interference is at the dimerisation interface.

The strategy we employed was to target not only the active site residues of TIM, but also allow ligands having contacts surrounding the active site region. These ligands might possibly have higher selectivity, because TIM is less conserved away from the active site. The ligands would then still have to be capable of providing steric hindrance to

substrate binding.

The two inhibitory compounds Direct Red and Direct Violet are both brilliantly coloured dyes used in the textile industry. They have been used in a recent study (Gao *et al.*, 1998) regarding the effect of Suramin and Suramin-like compounds on *Trypanosoma* TIM. The structures of Suramin and its similar compounds Direct Red and Direct Violet are shown in Figure 4.9, 4.10.

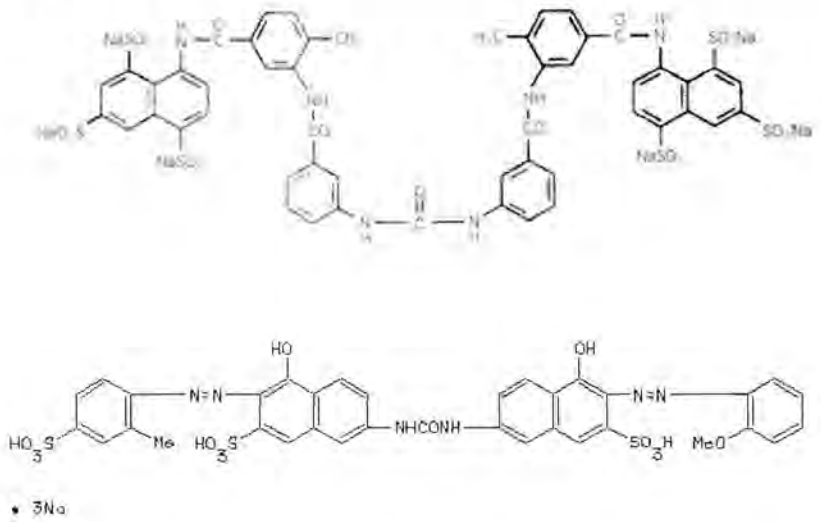


Figure 4.9: Structures of Suramin (top), and Direct Red (bottom).

Suramin is a negatively charged molecule that has been used extensively in the treatment of African sleeping sickness caused by *T. brucei brucei*, but is ineffective in Chagas disease with *T. cruzi* as causative agent. It is known to affect various enzymes including some in the glycolytic pathway of *T. brucei*, and has been found to inhibit the reactivation and dimer formation of *Trypanosoma* and human TIM after unfolding (Gao *et al.*, 1998). TIM is inactive in monomeric form, and the authors proposed that Suramin interfered with TIM dimerisation thus decreasing TIM activity. However, TIM was also shown to be inhibited by Suramin (Misset and Opperdoes, 1984) at concentrations lower than those used for inhibition of refolding. Suramin has been shown to compete with substrates (K_i = 0.1mM), indicating an influence at the active site (Lambeir *et al.*, 1987). This data indicates that it probably binds to TIM in a conformation that affects both the active site and the

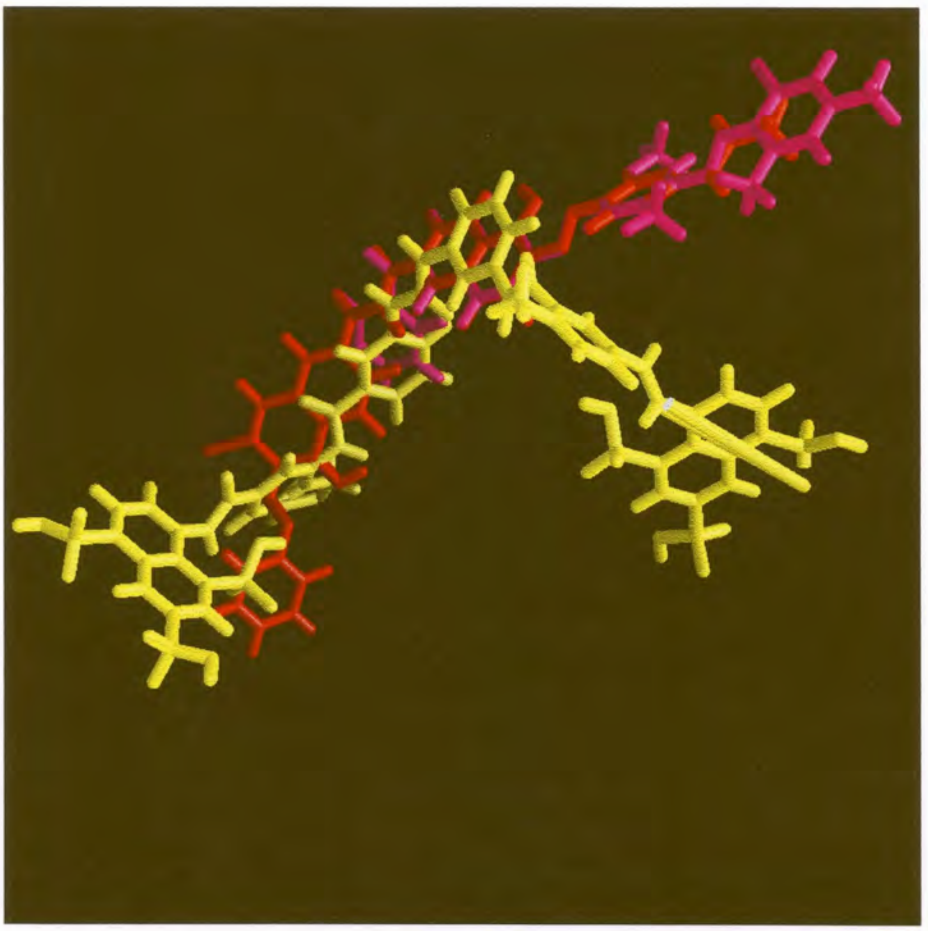
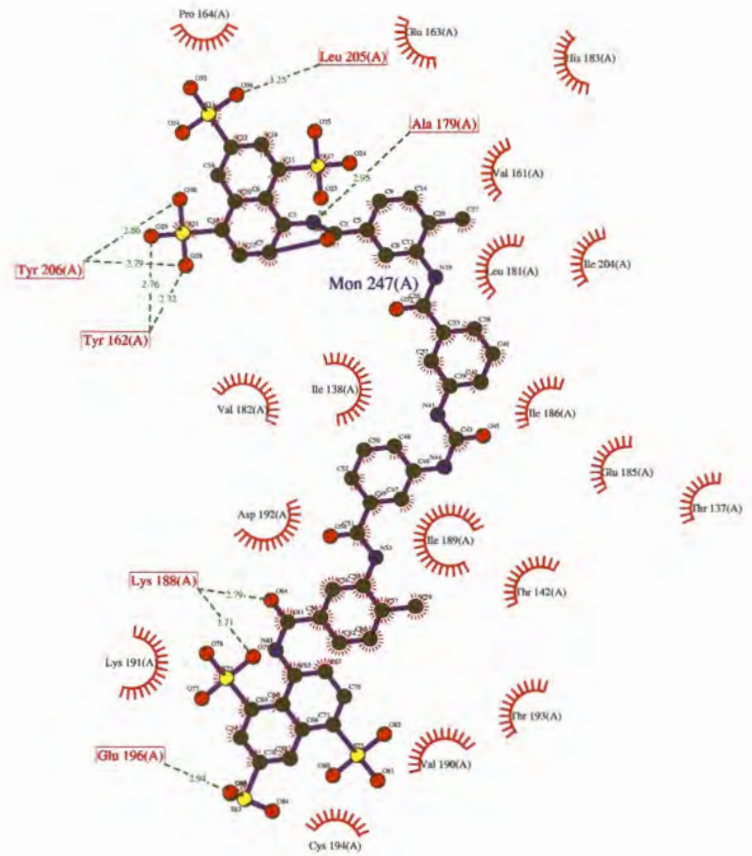


Figure 4.10: A comparison of the three-dimensional structures of Suramin (yellow), Direct Red (red) and Direct Violet (magenta).

dimerisation interface.

The two inhibitors Direct Red and Direct Violet were selected for TIM binding by the DOCK package, totally independent from the article regarding the effect of these compounds on *Trypanosoma* and human TIM (Suramin was not present in the ligand library). The binding scores of these two compounds together with the independent data from Gao *et al.* (1998) suggest that their binding occurs specifically to TIM. This counteracts previous speculations that the promiscuous interactions of Suramin with many different enzymes is simply due to the negative charge of the molecule. Although Suramin can be docked successfully into the TIM active site, no significant contacts of Suramin with the catalytic residues could be indicated (Figure 4.11).



Key

Ligand bond
Non-ligand bond
Hydrogen bond and its length
His 53 Non-ligand residues involved in hydrophobic contact(s)
Corresponding atoms involved in hydrophobic contact(s)

Suramin

Figure 4.11: Detailed contact map for the TIM-Suramin complex.

Mutational analyses of residues in contact with the two dyes may be used to confirm their conformation of binding to TIM. Ideally, the minimal fragment of these dyes capable of inhibiting TIM activity should be determined, and possibly used as a lead compound for anti-malarial

drug design. Lead compounds may be designed based on the common structural fragment as shown in Figure 4.8. Databases may initially be searched for molecules similar to the common fragment of both dyes, and these compounds then tested *in vitro*. Alternatively, packages are available for the *de novo* design of ligands based on a site contact map.

A comparison of the active sites for *Trypanosoma*, malaria and human TIM regarding residues having contact with Direct Red show little differences (Figure 4.12). Except for the mutation of Ser96 to Phe in the case of malaria, few other differences were visible. Both Direct Red and Direct Violet show contact with this residue, and it may be exploited for the design of species specific inhibitors.

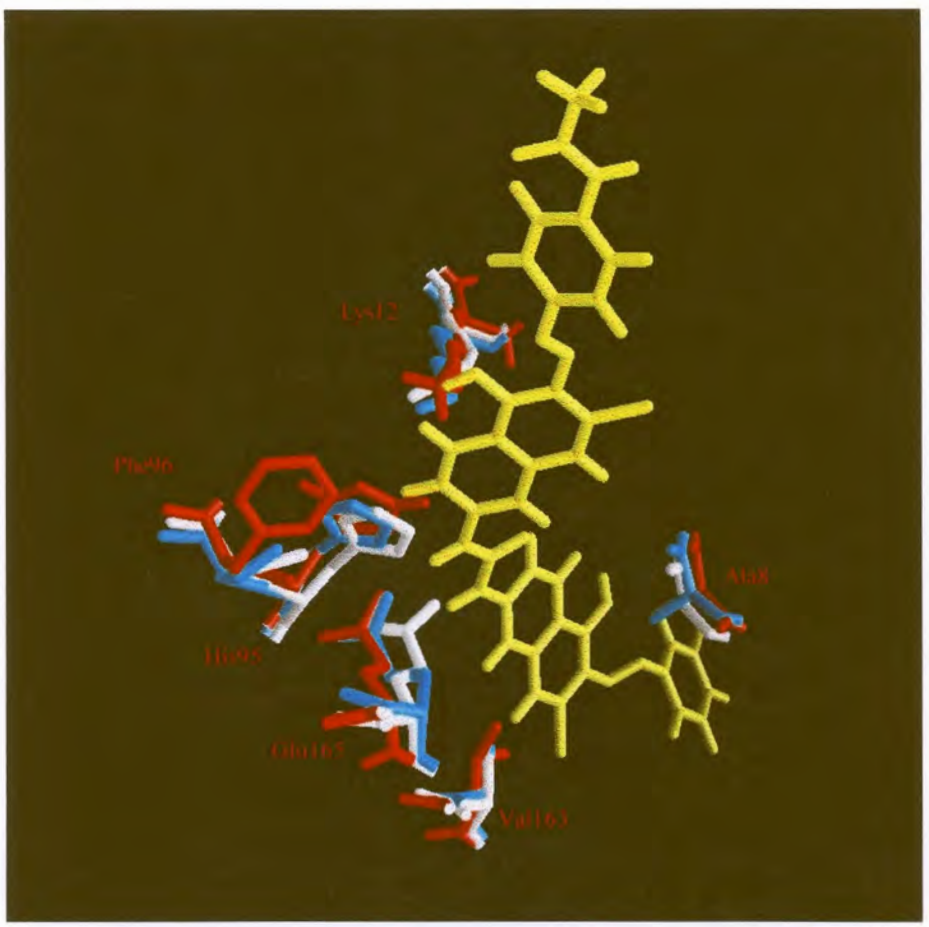


Figure 4.12: Active site comparison for malaria TIM (red), *Trypanosoma* TIM (cyan) and human TIM (white). Direct Red as docked into the malaria TIM active site is shown in yellow.

The mixed type inhibition that was found for Direct Red and Direct

Violet is probably due to a similar binding conformation to that of Suramin affecting both TIM dimerisation and the active site region. Docking studies of Direct Red and Violet do not show direct contact with the dimerisation interface of dimeric TIM, and it is suggested that the inhibitor binding probably causes perturbation of the TIM backbone to affect the dimerisation process.

Chapter 5

Concluding Discussion

The ultimate aim of this study was the generation of novel inhibitors against malaria metabolic enzymes. DHFR was chosen as candidate, since it has been validated as an efficient therapeutic target, and had been characterised in detail in various organisms. TIM has not been used as a drug target enzyme up to now, but plays a key role in the parasite's carbohydrate metabolism and has been studied extensively in many organisms.

To be able to design enzyme inhibitors, detailed knowledge is needed regarding the structure of the protein target. Unfortunately, in the case of malaria, only five enzymes have been crystallised. These are TIM (Velanker *et al.*, 1997), HGPRT (Shi *et al.*, 1999), LDH (Dunn *et al.*, 1996), fructose-1,6-bisphosphate aldolase (Kim *et al.*, 1998), and plasmepsin II (Bernstein *et al.*, 1999). DHFR has been cloned and crystallised from many organisms, and generally shows a high degree of amino acid sequence homology between species. Malaria DHFR was first expressed by Sirawaraporn *et al.* (1990). They were able to express low amounts of the DHFR-TS complex in *E. coli*. Speculating that the AT-rich malaria codons were inhibiting overexpression, a synthetic gene incorporating *E. coli* codon preferences was constructed. When overexpression of this DHFR domain was attempted, the protein was found only in inclusion bodies (Sirawaraporn *et al.*, 1993). A synthetic gene of DHFR-TS was recently prepared, which overexpressed in *E. coli* as an active correctly folded enzyme (Prapunwattana *et al.*, 1996). Up to now, this enzyme has not been crystallised. Initially, we set out to obtain DHFR in an active, correctly folded form. The malaria DHFR gene could be expressed using a T7 promoter, but no soluble expression of the DHFR domain

could be achieved despite the presence of solubility-enhancing fusion peptides. This study showed that a synthetic gene was not necessary for the overexpression of AT-rich genes, and that extremely high levels could be expressed under a T7-lac promoter. It was concluded that the presence of the TS-domain was crucial for the correct folding of malaria DHFR.

Homology modelling of malaria DHFR was complicated by the large insertions present in the DHFR gene compared to that of other species. These insertions could not be modelled from segments of other proteins and seem to be unique to malaria DHFR. Similar insertions have been found in various other malaria proteins, but the function for the inserted regions is unknown (LeBlanc and Wilson, 1993). Homology modelling of the active site was performed, but low quality scores did not yield sufficient confidence in the model to use it in ligand docking studies. Two recent articles discussed the use of homology modelled DHFR active sites in drug discovery and mechanisms of resistance. Lemcke *et al.* (1999) suggested the inserted regions occur as loops pointing away from the surface of the proteins. However, they could not conclusively offer proof regarding the loop conformations, and found relatively low quality values for these regions. They used their model to propose a mechanism for the influence of the *Ser108Asn* point mutation on pyrimethamine resistance, but could not yet explain the mechanism of the *Asn51Ile* mutation. Toyoda *et al.* (1997) prepared a homology-based model of DHFR by methods not yet published, and this was used for ligand design purposes. They were able to prepare inhibitors with micro- to nanomolar affinities in this way.

TIM has been extensively studied due to its role in diseases, as well as its structural properties. Although sequence identity is not high, the TIM fold is conserved in the enzyme from all species, and catalytic residues are conserved. Malaria TIM was first cloned by Ranie *et al.* (1993), and expressed in a TIM deficient *E. coli* strain. TIM activity was shown, but no isolations were performed. We obtained the TIM cDNA from local parasite isolates, and expressed TIM in the pET15b system, which supplies a N-terminal His-Tag for purification purposes. TIM was expressed in active form, and could be purified to homogeneity by immobilised metal affinity chromatography.

Due to the highly conserved nature of the TIM backbone, homology modelling was not as problematic as in the case of DHFR. The modelled backbone of malaria TIM showed very little deviation from that of

other species. The model was subsequently compared to the recently published X-ray structure of malaria TIM (Velanker *et al.*, 1997). In the active site, the only difference was the rotation angle of *Phe96*. This indicates the value of homology models in the drug design process when numerous templates for a protein with a conserved fold are available. Due to the discrepancy in the *Phe96* side-chain angle, the X-ray structure 1YDV was used in the rest of the study for ligand docking studies.

A ligand search was performed only for TIM, as DHFR could not be expressed in a soluble form and the DHFR homology models were not of sufficient quality. TIM is active only in dimeric form, thus lead compounds may be targeted againt the dimerisation interface as well as the active site. Our approach was to prepare a grid of the active site, whereupon a series of positions of each ligand could be scored. This approach minimises computational time by evaluating the protein chacteristics at each grid position only once. A set of spheres was then constructed to limit the ligand binding to the active site area. The NCI-3D database was screened against the binding site with bump filtering to eliminate molecules which would not fit, before scoring took place. Ideally, a chemical library such as the ACD should be used during screening, as all ligands are commercially available. Many of the NCI-3D compounds have been custom synthesized, and are extremely difficult to obtain. It should also be taken into account that the 3D-version of the NCI database has been constructed by means of the Corina packages, and the theoretically optimal conformation is not always the major isomere found during synthesis. For example, the highest ranking ligand, bicyclo[2.2.1]heptane-2,3,5,6-tetracarboxylic acid, is found in an endo-endo conformation in the NCI library, but is very difficult to synthesise in that conformation.

The 100 top scoring compounds were inspected, and duplicates eliminated which differed only in bound salts, and some similar substituents. Water insoluble compounds were also eliminated together with highly toxic or carcinogenic compounds. Of the remaining compounds, 7 were purchased commercially and 5 were chosen for synthesis. When tested for *in vitro* inhibition of TIM activity, the compounds Direct Red 23 and Direct Violet 51 (hereafter referred to as Direct Red and Direct Violet) which are commercial dyes, were the only compounds showing a decrease in activity in the $<100\mu$M range. Upon literature investigation of these and related compounds, it was found that a recent study had been performed by Gao *et al.* (1998), investigating the effect of these

compounds on reactivation of *Trypanosoma* and human TIM. The primary compound used in this study was Suramin, a drug used for sleeping sickness which was originally based on the trypanocidal effect of Trypan Red. Suramin is known to also affect a series of other enymes in the glycolytic pathway (Brunner *et al.*, 1996)(Gonzalez and Cazzulo, 1989)(Misset *et al.*, 1986) as well as tumor necrosis factor α (Alzani *et al.*, 1993). It was speculated that Suramin interfered either with TIM monomer folding or with monomer association. Although Suramin has been shown to inhibit TIM (Misset and Opperdoes, 1984), refolding was affected at concentrations $<1\mu$M, a concentration shown not to affect TIM activity. Preincubation with Suramin followed by dilution did not affect TIM activity. It was also found that the effect on TIM renaturation was protein concentration dependent, at concentrations above 1μg/ml the effect vanished. It was suggested that Suramin acted by the formation of Suramin-mTIM complexes which failed to dimerise, and occured only when the monomers had existed for a substantial time. When testing the related compound Direct Red, a 10x more potent inhibitory effect was found. They could show no common structural feature between all the inhibitors.

Our results indicate that Direct Red and Direct Violet do in fact bind in the active site of TIM, leading to uncompetitive inhibition of TIM activity. This suggests that these compounds may bind to the active site in such a way as to also disrupt the overall TIM conformation, thus interfering with successful dimerisation. A common TIM-bound three-dimensional structural motif could be shown for Direct Red and Direct Violet (Figure 4.8), and a similar region occurs on the other compounds shown to affect TIM reactivation, with Suramin and Direct Yellow showing the least similarity as well as the lowest activity (Figure 5.1). It is hoped that this may be the basis for the design of species-specific inhibitors.

Once the synthetic compounds have been screened, those showing inhibition of TIM activity will also be inspected for a structural motif similar to the one indicated above. This motif may then be consolidated into the minimal fragment capable of inhibiting TM activity. Chemical searches may be performed to find the existing compounds with structures closest to the desired motif. These may then be chemically modified to optimise a TIM inhibitor showing the highest inhibition of malaria TIM and the smallest effect on human TIM.

There are two basic metabolic methods for killing an organism, either

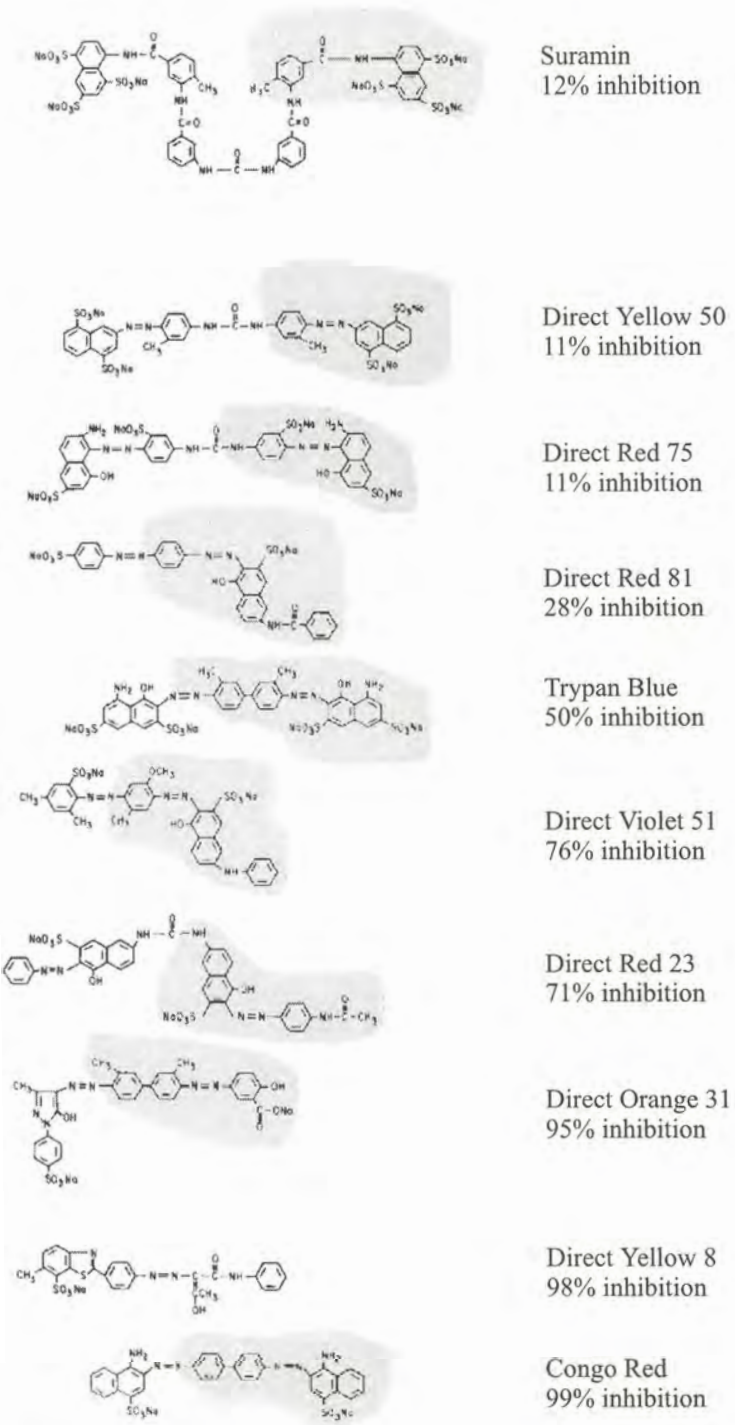


Figure 5.1: Structures of Suramin and related compounds. Percentage inhibition of *Trypanosoma* TIM reactivation at 10μM is indicated. Similar structural motifs are shaded in grey (Gao *et al.*, 1998).

the flux through an essential metabolic pathway may be decreased, or a metabolite concentration can be increased to toxic levels. Decrease of flux will normally require a tight-binding inhibitor with a significant flux control coefficient, while an uncompetitive inhibitor with a small flux control coefficient will lead to the increase of metabolites to toxic levels (Eisenthal and Cornish-Bowden, 1974). Thus, uncompetitive-type inhibitors are also of value as lead drug compounds. Of the glycolytic enzymes in *Trypanosoma*, Bakker *et al.* (1999) identified the glucose transporter to be the most important, followed by fructose-1.6-bisphosphate aldolase (ALD), glycerol-3-phosphate dehydrogenase (GDH), glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and phosphoglycerate kinase (PGK). In red blood cells, deficiencies of ALD, GAPDH and PGK did not cause any clinical symptoms. It has been shown that no organism can survive with a serious TIM deficiency, making it a good therapeutical target, but also indicating the dangers of TIM inhibitors for the human patient.

With the need for novel malaria drugs increasing on a daily basis, all studies showing promise of lead drugs should be followed up to as great an extent as possible. As drug companies are generally not interested in potential therapeutical targets until the *in vivo* value has been proven, compounds with a history of clinical applications such as the dyes described in this thesis should be regarded as serious contenders for the design of clinically useful agents.