Chapter 1

Introduction

1.1 Polyamines

The polyamines putrescine, spermine and spermidine (Fig. 1.1) are near ubiquitous polycationic aliphatic amines required for a number of essential cellular processes, particularly in organisms undergoing rapid proliferation. Putrescine is typically formed from ornithine by ornithine decarboxylase (ODC) which then serves as a scaffold for the addition of amino-propyl groups from decarboxylated *S*-adenosylmethionine (formed by *S*-adenosylmethionine decarboxylase: AdoMetDC) to produce spermidine (spermidine synthase) and spermine (spermine synthase), respectively (Fig. 1.2). Spermidine and spermine can also be back-converted to their precursors via the combined action of spermine/spermidine N^1-acetyltransferase (SSAT) and polyamine oxidase (PAO). Ornithine is produced from arginine by arginase to release urea. Alternatively, arginine may also serve as the source of putrescine via arginine decarboxylase and agmatine ureohydrolase (Fig. 1.2, Tabor and Tabor, 1984, 1985; Cohen, 1998).

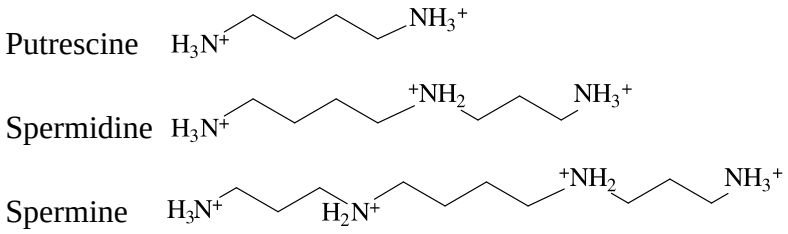


Figure 1.1: The polyamines: putrescine, spermidine and spermine.

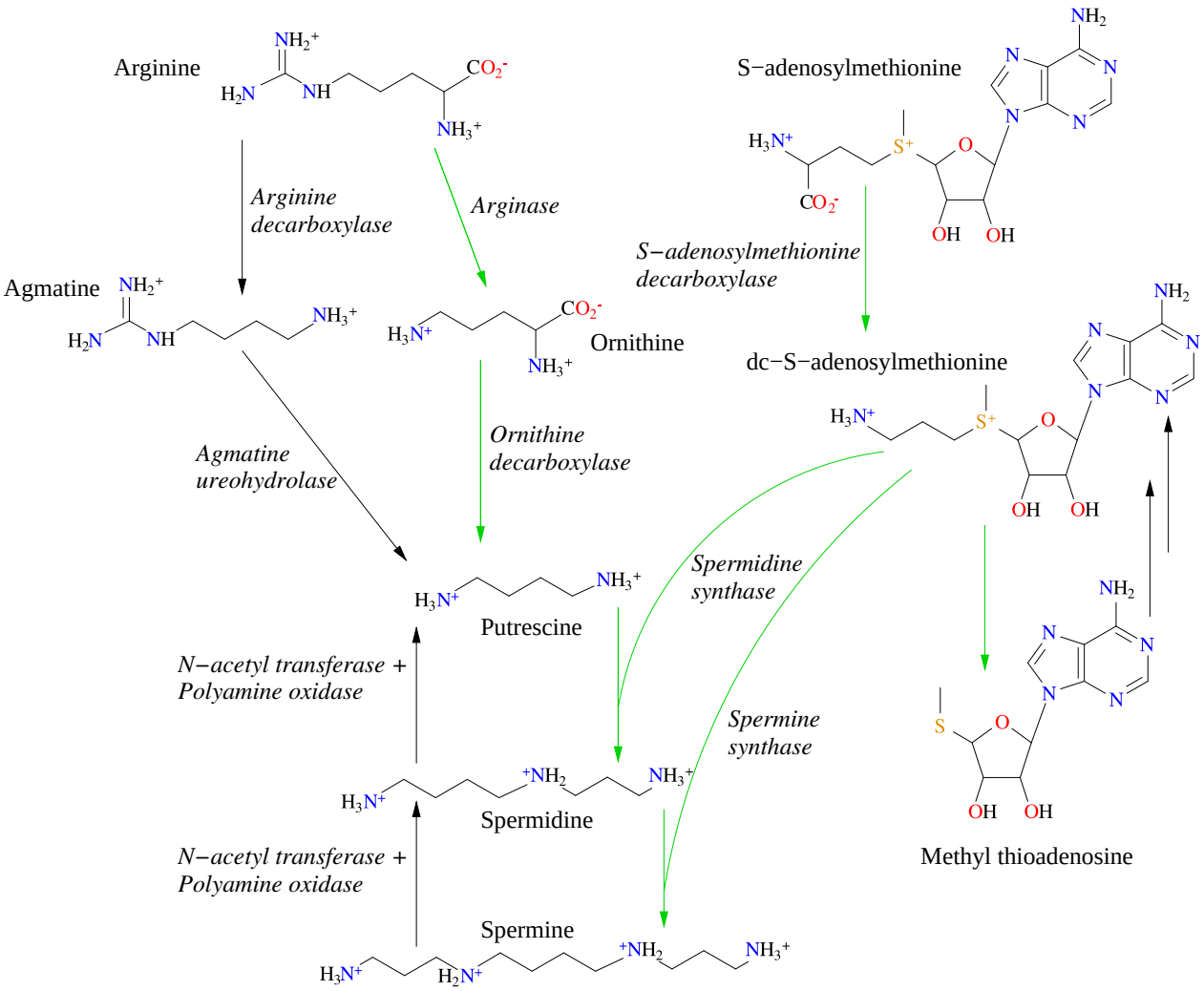


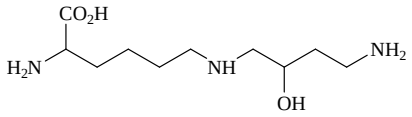
Figure 1.2: Outline of polyamine metabolism. Pathways that have been identified in the malaria parasite *Plasmodium falciparum* are indicated in green. The inclusion of spermine synthase is by virtue of low levels of spermine synthesis by spermidine synthase.

The polycationic nature of polyamines enables them to interact electrostatically with large biological macromolecules such as DNA/RNA and proteins (Bachrach, 2005). It has been suggested that within the nucleus polyamines form aggregates mediated by phosphate ions. These so-called nuclear aggregates of polyamines (NAPs) in turn interact with DNA (D'Agostino and Luccia, 2002; D'Agostino *et al.*, 2005; Luccia *et al.*, 2009). Polyamines can thus affect DNA conformation and chromatin remodeling by enhancing DNA condensation within the tight confines of the nucleus. This in turn affects DNA stability and transcription (Childs *et al.*, 2003; Wallace *et al.*, 2003; Janne *et al.*, 2004).

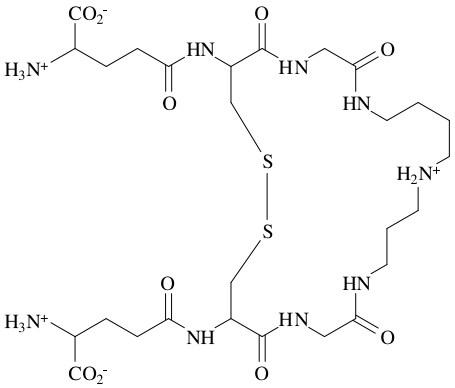
Polyamines are also known to interact with various proteins with varying effects depending on the polyamine species, concentration and protein species. Casein kinase 2 (CK2) interacts with spermine via it's β regulatory subunit leading to enhanced activity. While the biological function of CK2 remains uncertain it has been linked to malignancy via one of its substrates, the oncoprotein Myc. Spermine is also known to modulate the function of membrane proteins such as *N*-methyl-D-aspartate (NMDA) receptors. The formation of protein-DNA complexes is also affected by the presence of polyamines. Complex formation is

typically enhanced, although inhibition may be observed at high polyamine concentrations. Polyamines have also been observed to affect protein degradation, depending on concentration and protein species (Childs *et al.*, 2003).

Additionally polyamines are also required to form certain secondary metabolites, such as the post-translationally modified amino acid hypusine (Tabor and Tabor, 1984; Park *et al.*, 1993) and of the glutathione analogue, trypanothione, in *Trypanosoma* (Fig. 1.3 Müller *et al.* 2003; Heby *et al.* 2007). Hypusine is a post-translationally modified amino acid formed from specific lysine residue of the eukaryotic initiation factor 5A (eIF5A). Firstly, butylamine is transferred from spermidine to the side-chain amino group by deoxyhypusine synthase. This is followed by β-hydroxylation (deoxyhypusine hydroxylase) to form hypusine (Park *et al.*, 1981, 1982). In the protozoan parasites of the trypanosomatid family spermidine is conjugated with glutathione to form trypanothione which replaces the usual glutathione redox system. Glutathione is first conjugated with spermidine to produce glutathionylspermidine by glutathionylspermidine synthetase, followed by a further glutathione conjugation to form trypanothione (Müller *et al.*, 2003; Heby *et al.*, 2007).



(a) Hypusine



(b) Trypanothione

Figure 1.3: Secondary metabolites formed from spermidine

Polyamines also bind to various species of RNA. Binding stabilises tRNA and affects the conformation of 16s rRNA (Amarantos and Kalpaxis, 2000; Amarantos *et al.*, 2002). Furthermore, the presence or absence of polyamines causes translational frame-shifting in certain mRNAs. The Ty1 transposable element in yeast undergoes an increased +1 frame-shift to produce the TYA-TYB fusion protein. During polyamine depletion there is increased +1 frame-shifting and transposition (Clare *et al.*, 1988). In many eukaryotes ODC is regulated by antizyme (AZ) which binds to ODC and targets it for non-ubiquitin mediated proteolysis. AZ mRNA comprises two overlapping open reading frames. In the presence of polyamines the ribosome undergoes a frame-shift to produce the functional protein (Rom and Kahana, 1994).

Tight regulation of polyamines is required for progression through the cell cycle (Ackermann *et al.*, 2003). During the G1 and G2 phases an increase in cellular polyamines is generally observed. Inhibition of polyamine biosynthesis is often observed to arrest cell growth. Furthermore, polyamines are also associated with apoptosis. The association remains uncertain however, with both increased and decreased polyamine levels being linked to both increased and decreased apoptosis. (Wallace *et al.*, 2003).

Polyamine biosynthesis has been identified as a possible therapeutic target for various parasitic diseases (Müller *et al.*, 2003; Heby *et al.*, 2007), cancers (Wallace, 2007) and even HIV via the requirement for hypusine (Schafer *et al.*, 2006). Polyamine biosynthesis enzymes characterised in the malaria parasite *P. falciparum* include the bifunctional *S*-adenosylmethionine decarboxylase/ornithine decarboxylase (Müller *et al.*, 2000; Krause *et al.*, 2000; Wrenger *et al.*, 2001; Birkholtz *et al.*, 2003, 2004), spermidine synthase (Haider *et al.*, 2005) and arginase (Müller *et al.*, 2005).

1.2 Malaria

1.2.1 Introduction and prevalence

Malaria is caused by protozoan parasites of the *Plasmodium* genus and is transmitted by the female *Anopheles* mosquito. Five species of *Plasmodium* are known to infect humans: *P. falciparum*, *P. vivax*, *P. ovale*, *P. malariae* and *P. knowlesi* (which till recently was thought to only infect macaques, Singh *et al.* 2004). Of these *P. falciparum* is the most virulent, causing the most deaths. *P. vivax* is the second most dangerous but is only common in tropical regions outside of Africa (Fig. 1.4), the continent with the largest malaria burden (Hyde, 2007; Greenwood *et al.*, 2008). Currently about 2 billion people are at risk of malaria resulting in about 500 million cases annually and 1 million deaths. The majority of the burden exists in developing countries of the tropics and sub-tropics with the majority of casualties being among children. Malaria thus represents an significant impediment to the economic development for much of the world. (Snow *et al.*, 2005; Rowe *et al.*, 2006; Greenwood *et al.*, 2008; Hay *et al.*, 2009).

Plasmodium parasites exhibit a complex life cycle involving a vertebrate host and an invertebrate host (Fig. 1.5). Infection of the vertebrate host begins with inoculation with sporozoites by a mosquito of the *Anopheles* genus. The sporozoites then travel via the blood stream to the liver where they infect hepatocytes. During this asymptomatic stage the sporozoites multiply asexually, eventually releasing themselves from the hepatocyte as merozoites. The merozoites then in turn infect erythrocytes for further rounds of asexual reproduction, passing through various stages (ring → trophozoite → schizont), eventually bursting the red-blood cell to release further merozoites or gametocytes. Free gametocytes are then taken up by another *Anopheles* mosquito, where sexual reproduction occurs (Greenwood *et al.*, 2008).

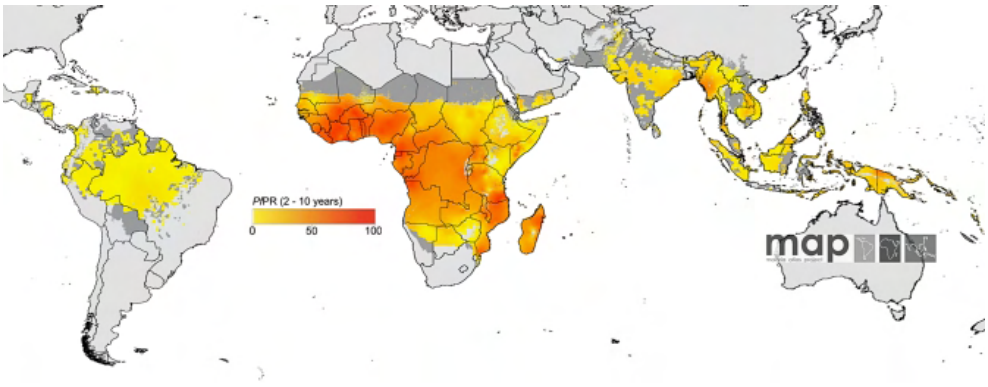


Figure 1.4: Endemicity of *P. falciparum* for 2007 measure as the *P. falciparum* parasite rate (*PfPR*, percentage population with detectable levels of parasites in the blood) for the 2 - 10 year old age group. Adapted from Hay *et al.* (2009).

Global attempts to eradicate malaria beginning in the 1950s achieved partial success in some parts of the world outside of Africa. The two main components of this campaign were the use of chloroquine for treatment and dichloro-diphenyl-trichloroethane (DDT) for vector control. However resistance to both these interventions evolved and the campaign was never attempted in Africa, where there is the highest intensity of malaria transmission. Furthermore, resistance has evolved towards sulphadoxine-pyrimethamine, the front-line treatment that replaced chloroquine (Greenwood *et al.*, 2008).

Current anti-malarials target a number of cellular processes, mostly within the asexual erythrocytic stages of the parasite. Within the cytosol, folate biosynthesis is targeted via inhibition of dihydrofolate reductase (DHFR) and dihydropteroate synthase (DHPS). Antifolates remains the most common anti-malarial drug class, including pyrimethamine, proguanil, dapsone and sulphadoxine. A DHFR inhibitor is typically used in combination with a DHPS inhibitor, e.g. pyrimethamine and sulphadoxine (fansidar) or proguanil and dapsone. The quinoline family of drugs including quinine, chloroquine, amodiaquine, mefloquine, halofantrine and lumefantrine sequester in the digestive food vacuole of the erythrocytic stages. While their mechanism remains generally unknown it is likely to be mediated by binding haem thus inhibiting haem detoxification. A number of anti-bacterials target the parasite by inhibiting translation within the apicoplast, a chloroplast derived organelle. These include azithromycin, clindamycin and doxycycline. Fosmidomycin acts on the isoprenoid biosynthesis pathway within the apicoplast. Lately artemisinin and it's derivatives have received a lot attention for it's ability as a fast acting drug. Artemisinin is typically combined with longer acting drugs and is notable in that it also targets gametocytes, thus reducing transmission. The mechanism of artemisinin's action remains a topic of considerable debate (Hardman and Limbird, 2001; Greenwood *et al.*, 2008). Resistance has been detected in most drugs in current use (Hyde, 2007). Resistance to artemesinin has been observed in the lab and recent reports indicate this has also emerged in the field on the Thailand-Cambodia border (Dondorp *et al.*, 2009).

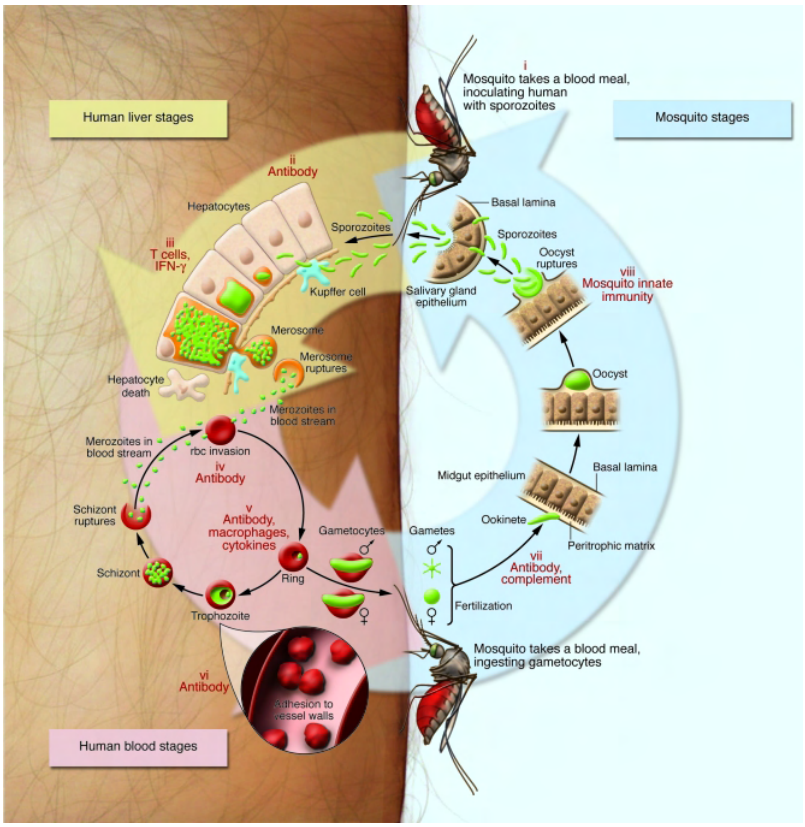


Figure 1.5: Life cycle of *Plasmodium*. Sporozoites are inoculated by a female *Anopheles* mosquito directly into the blood or more often into the dermis (and must then travel to the circulatory system). Upon reaching the liver the sporozoites infect hepatocytes, which rupture and release merozoites approximately one week later. The merozoites then infect erythrocytes to initiate the red-blood cell stages. The parasite then passes through the ring, schizont and trophozoite stages, later rupturing the erythrocyte to release yet more merozoites for further infection. Alternatively, the bloods-stages may develop into the gametocytes to be taken up by an *Anopheles* mosquito during the next blood meal. Adapted from Greenwood *et al.* (2008).

A number of new targets have been identified in recent years, especially as a result of the *Plasmodium* genome sequencing projects. Among these are proteases specific to the digestive food vacuole required for haemoglobin degradation, fatty acid and isoprenoid biosynthesis within the apicoplast, the shikimate pathway as well as lactate, orotate and inositol metabolism (Gardner *et al.*, 2002).

Attempts to generate a malaria vaccine have resulted in mixed success. Vaccination with radiation attenuated sporozoites and sporozoite derived subunits provides partial protection but with waning efficacy. Attempts to immunise against the erythrocytic stages have yet to demonstrate significant protection. In non-human animals (Darwin, 1859) immunisation against the sexual stages has been demonstrated to prevent transmission. While this does not reduce disease it might prove useful in reducing transmission in humans (Greenwood *et al.*, 2008).

Apart from chemotherapy, vector control constitutes the other major arm of malaria eradication. Distribution of insecticide treated nets (ITN) increase survivability in children,

while indoor residual spraying (IRS) remains effective in certain areas. However, resistance to pyrethroids, the insecticide most used in ITNs is increasing and DDT only remains effective in restricted areas (Greenwood *et al.*, 2008).

There is thus an urgent need for new anti-malarial strategies if this global problem is to be dealt with.

1.2.2 Polyamine metabolism as a *Plasmodium* drug target

Polyamines are essential for cell growth, proliferation and differentiation. Because of this their metabolism has received a lot of attention as a possible drug target, especially within the cancer research community. Consequently a number of potent inhibitors of polyamine metabolism enzymes have been discovered and developed over the years (Fig. 1.6). Attempts to use these to target polyamines in anti-cancer therapy have largely been disappointing, however. In general, inhibiting polyamine biosynthesis induces cytostasis instead of cytotoxicity within tumour cells (Marton and Pegg, 1995). The reason for the lack of anti-tumour effects is largely due to compensatory mechanisms within the mammalian cell for maintaining the polyamine pools. It has been identified that multiple components of polyamine regulation will have to be targeted when targeting this metabolism for cancer (Seiler, 2003a). Specifically, the uptake of exogenous polyamines by transport systems will likely have to be targeted along with polyamine biosynthesis (Seiler, 2003b). Among the inhibitors identified, the most well known is alpha-difluoromethylornithine (DFMO). DFMO has been used successfully to treat West African Sleeping sickness caused by *T. brucei gambiense* (Wang, 1995). The low toxicity of this drug and its ability to penetrate the blood-brain barrier contribute to it's effectiveness in this regard. The dependence of *Trypanosomes* on the spermidine derived glutathione analogue trypanothione renders them particularly vulnerable to polyamine depletion. The viability of polyamine metabolism as a malarial drug target has been extensively reviewed by Müller *et al.* (2008). DFMO has been found to inhibit *P. berghei* sporogony in *Anopholes stephensi* (Gillet *et al.*, 1983) as well as protect mice from sporozoite infection (Lowa *et al.*, 1986), suggesting that DFMO may have utility as a prophylactic. Targeting the blood stages has proved less promising, however. Inhibition of *Pf*ODC decreases putrescine levels (Wright *et al.*, 1991), while *Pf*AdoMetDC inhibition reduces both spermidine and spermine (Gupta *et al.*, 2005). Cytostasis at the trophozoite stage can be induced by inhibition of *Pf*ODC or *Pf*AdoMetDC but mice infected with *P. berghei* are not cured (Assaraf *et al.*, 1987; Bitonti *et al.*, 1987; Gupta *et al.*, 2005). Co-inhibition of *P. falciparum* ODC/AdoMetDC induces partial transcriptional arrest at the trophozoite stage and transcriptional regulation of proteins in response to polyamine depletion (van Brummelen *et al.*, 2009). Arrest at the trophozoite stage can be rescued by the addition of exogenous putrescine (Assaraf *et al.*, 1987), however, the effect of exogenous spermidine and spermine remains uncertain due to conflicting reports (Assaraf *et al.*, 1987; Wright *et al.*, 1991; Bitonti *et al.*, 1987). Exogenous putrescine is observed to accumulate within the parasites of infected erythrocytes, while erythrocyte levels remain similar to unin-

fected cells. Uptake of exogenous spermidine and spermine has been suggested by the work of Fukumoto and Byus (1996) but also remains uncertain due to conflicting reports (Gupta *et al.*, 2005). Despite these conflicting results, inhibition of polyamine uptake is considered an important target within the parasite. Other polyamine metabolism and related enzymes that have been identified as potential drug targets include *S*-adenosylmethionine synthetase (that produces AdoMet), spermidine synthetase and methylthioadensine (the product of AdoMetDC) recycling enzymes (Müller *et al.*, 2008). To date the most promising result has been obtained when both *Pf*ODC and *Pf*AdoMetDC are inhibited together with exogenous polyamine uptake by the inclusion of bis(benzyl)polyamine analogues. This combined approach was found to be curative of *P. berghei* infected mice (Bitonti *et al.*, 1989). Unpublished results of Brun and Walter with the AdoMetDC inhibitor CGP 40215A produced curative results in *P. berghei* infected mice but without affecting polyamine levels, thus the target remains unclear (Müller *et al.*, 2008). While current results are mixed, polyamine metabolism within *Plasmodium* remains a target worth investigating.

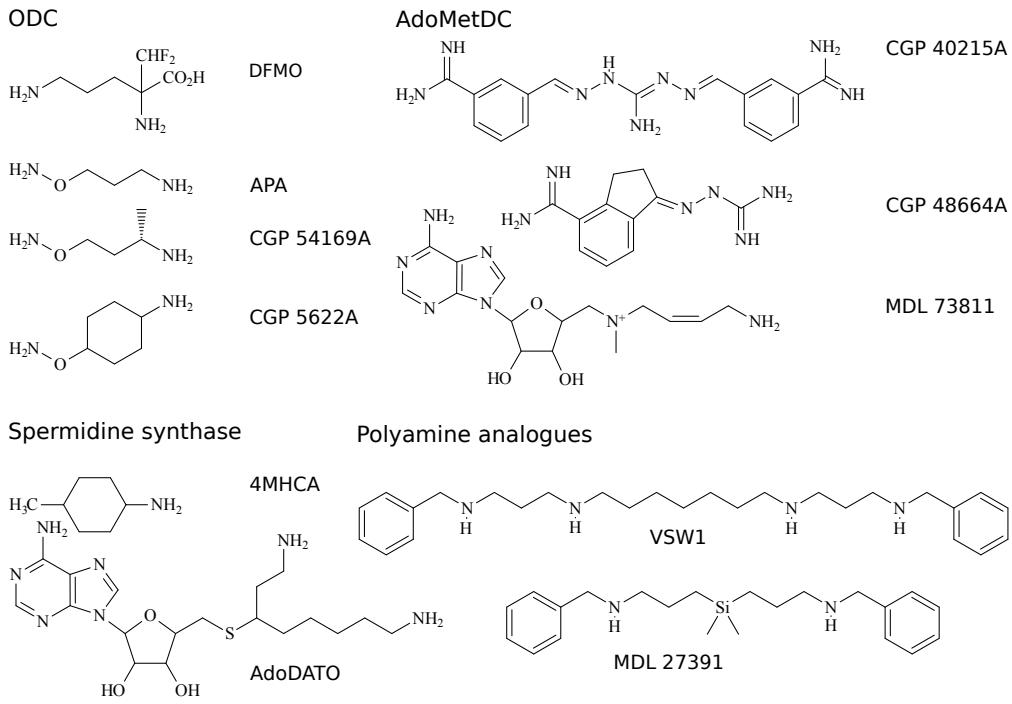


Figure 1.6: Common inhibitors of polyamine metabolism.

1.3 Computational structural biology and rational drug design

1.3.1 Rational drug design

High throughput techniques in modern drug chemistry allow for the generation of large libraries in the order of thousands to millions of compounds. This can be similarly followed up by high throughput screening assays to identify novel lead compounds that may ultimately

become or serve as scaffolds for new drugs. Following this approach blindly has only produced modest results, however. Successful high-throughput screening depends on striking a good balance between structural diversity of the compound library while not over-sampling futile regions of chemical space (Lipinski *et al.*, 2001; Snowden and Green, 2008). The average cost of bringing a new drug to market is in the order of $800-900 million (DiMasi *et al.*, 2003; Vernon *et al.*, 2009). Much of the cost results from attrition of lead compounds during the late stages of the research pipeline due to problems with biological availability and toxicity.

A number of computational techniques have been developed in recent years that refine this process and have the potential to substantially reduce the cost of drug discovery. Rational drug discovery depends on having structural information of potential binding compounds and/or the protein target in question. Considering that many drugs act by binding an enzyme active site, the first question that often arises is whether ligand binding can occur for the target protein. A number of docking algorithms exist to predict ligand binding if the structure of both the protein and ligand are at hand. During protein docking various conformational combinations of the ligand and/or protein are sampled and scored in order to determine the most energetically favourable binding. Docking can be applied to whole compounds from large libraries numbering in the millions or to individual fragments which can later be assembled into a larger compound with higher affinity (Klebe, 2006; Orry *et al.*, 2006). Although accurate sampling is computationally expensive, such *in silico* screening is fortunately also embarrassingly parallel and can be distributed across many thousands of computers. While *in silico* docking is not yet absolutely reliable (Ferrara *et al.*, 2004) it can be used to screen out compounds that are unlikely to bind and focus on the most promising candidates.

When the protein structure is absent it remains possible to predict the binding of potential leads by extrapolating from the activities of known ligands and their structures. Using various statistical methods it is possible to extract quantitative structure activity relationships (QSAR) comprising 2D and/or 3D structural descriptors that contribute positively or negatively to activity. These QSARs can then similarly be used to screen the activity of novel compounds (Böhm *et al.*, 1999; Livingstone, 2000). An example of QSARs is given in Figure 1.7.

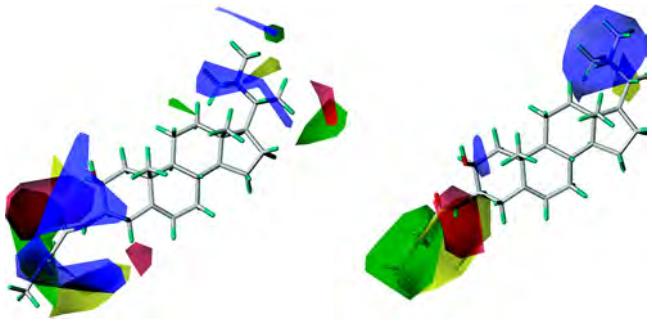


Figure 1.7: Examples of 3D QSARs derived from 39 butyrylcholinesterase inhibitors. A Comparative Molecular field Analysis (CoMFA) QSAR (left) and a Comparative Molecular Similarity Analysis (CoMSIA) QSAR (right). Sterically favourable (green), sterically unfavourable (yellow), positively charged favourable (blue) and positively charged unfavourable (red) regions are depicted. Adapted from Zaheer-ul *et al.* (2008).

In addition to predicting binding it is also important to understand a compound's effect at the organismal level. It is not sufficient for a compound to bind the target in question. A lead should also possess good properties with regards to absorption, distribution, metabolism, excretion and toxicity (ADMET). Unpromising drug leads can therefore be screened out using QSARs with respect to ADMET. Due to the high attrition rate of drug leads at the late stage in development it is becoming essential to screen out as many bad candidates early in the drug discovery process as possible. The most popular of these screens is "Lipinski's rule of 5", whereby it was observed that most sold drugs posses five or less hydrogen bond donors, 10 or less hydrogen bond acceptors, a molecular weight of 500 Da or less and a $c\log P$ of 5 or less (Lipinski *et al.*, 2001).

A good example of the application of rational drug discovery methods for malaria is the so-called World-Wide In Silico Docking of Malaria (WISDOM, http://wisdom.healthgrid.org) project. During the first round of WISDOM about 1 million compounds were docked using FLEXX (Rarey *et al.*, 1996) and AUTODOCK (Goodsell *et al.*, 1996) on two *P. falciparum* plasmepsins using a distributed network on the EGEE (Enabling Grids for E-sciencE, http://www.eu-egee.org) computing grid. A number of known inhibitors were identified as well as a class of novel guanidino based inhibitors from about 40 million dockings. Some if these compounds were later confirmed to be active *in vitro*. During round two dihydrofolate reductase from *P. falciparum* and *P. vivax* as well as glutathione-*S*-transferase from *P. falciparum* were docked against the same ZINC derived library using FLEXX to yield about 140 million dockings. In addition to EGEE a number of other European and affiliated grids were used (Kasam *et al.*, 2009).

1.3.2 Structural modeling

To fully exploit all the methods of rational drug design it is necessary to have the 3D structure of the target protein. The most reliable protein models are generated experimentally via

X-ray crystallography and NMR. Current protein structure determination methods are time consuming, however (Bourne and Weissig, 2003). The Protein Data Bank (PDB) currently contains $>$ 57 000 protein structures largely determined using X-ray diffraction and NMR (http://www.rcsb.org). In contrast the number of protein sequences grows much faster. For example just the well-curated SwissProt protein database contains $>$ 500 000 sequences (http://au.expasy.org/sprot/relnotes/relstat.html). Despite the efforts of structural genomics projects which have rapidly increased the pace of protein structure determination, there is a considerable gap between high-throughput structure and sequence data. Newer methods such as high-energy X-ray based methods under development promise to eventually allow for direct determination of protein structure in solution (Mardis *et al.*, 2009; Tiede *et al.*, 2009). Until such techniques become standard computational modeling can fill the gap. Furthermore, for reasons discussed below, *Plasmodium* proteins have proved more difficult than usual to crystallise. For these reasons it is often necessary to follow *in silico* based methods to determine the structures of *Plasmodium* proteins. In this study computational methods are relied on heavily to understand the unique structural features of certain proteins from *Plasmodium* polyamine biosynthesis metabolism.

The holy grail of structural modeling is to be able to predict the 3D structure from sequence alone. The theoretical basis of this is the assumption that the native protein fold is also the global energy minimum of the macromolecule. Therefore, with an accurate mathematical representation of the protein it should be possible to predict the structure by predicting the global minimum. A number of approaches are available for this. The most reliable method would be to model the molecule using computational quantum chemistry. Due to the large computational resources required, however, this approach can only be followed for small molecules (a few hundred atoms) and is generally still not feasible for molecules the size of proteins. Liu *et al.* (2001) have demonstrated with crambin that quantum mechanical simulation is possible using a high performance computing and a semi-empirical quantum representation. It is more common to represent only part of a protein structure quantum-mechanically, e.g. the active site. Although quantum mechanics provides the most accurate description, other methods are required for routine modeling (Leach, 2001; Schlick, 2002).

A more feasible approach is to represent a protein molecule using classical (Newtonian) mechanics. In this treatment the molecule is split up into geometrical components. The terms/components most commonly included are bonds, bond-angles, torsions, improper torsions, electrostatic interactions and Van der Waals interactions (Fig. 1.8). The energy of each component is included in a large sum describing the molecule that can be referred to as the scoring or energy function. The collection of mathematical forms that is used to describe each geometrical component is referred to as the force field. Force fields are typically designed to give a physical description of the molecule in question that could be computed in reasonable time. More complicated force fields include cross-terms to account for interactions between components. Terms can also be derived statistically from distributions of

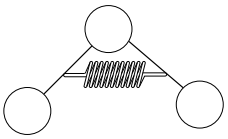
known protein structures and possibly combined with geometric components (Leach, 2001; Schlick, 2002).

$$V = \sum_{bonds} k_b(b - b_0)^2$$

Bonds

- b_0 - reference bond length
- b - bond length
- k_b- force constant

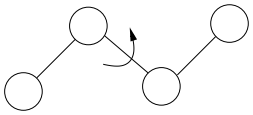
$$V = \sum_{angles} k_\theta(\theta - \theta_0)^2$$



Bond angles

- θ_0 - reference bond angle
- θ - bond angle
- k_θ- force constant

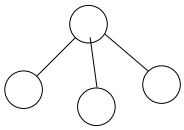
$$V = \sum_{dihedrals} k_\phi(1 - cos(n\phi - \delta))$$



Dihedral/torsion angles

- ϕ - Dihedral angle
- δ - Phase shift
- n - Periodicity
- k_ϕ- Dihedral force constant

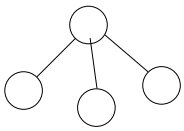
$$V = \sum_{impropers} k_\omega(\omega - \omega_0)^2$$



Improper torsions

- Planer components, e.g. aromatic rings
- $\omega - \omega_0$ - Out of plane angle
- k_ω- Force constant

$$V = \sum_{U-B} k_u(u - u)_0^2$$



Urey-Bradley terms

- Non-bonded atoms connected by 2 bonds:
- Pseudo-bond between 1st and 3rd atoms
- k_ω- Force constant

$$V = \sum_{non-bonded} \epsilon \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}}$$

Non-bonded interactions

- Atoms $>$ 4 bonds apart
- First term: Van der Waals
- Second term: Electrostatic
- r_{ij} - distance between atoms i and j

Figure 1.8: The components of a typical forcefield. Atomistic systems are modelled essentially as "balls on springs" using Newtonian mechanics.

Forcefields are typically used to find an energy minimum (minimisation) as well as for modeling the evolution of a protein in time (molecular dynamics). Using a combination of these approaches structure can be predicted *in silico*. Protein folding typically occurs on a microsecond to second time scale, whereas most molecular dynamics simulations are in the nano-second order. However, with recent advances in methods and hardware it is possible to model proteins *ab initio*. This is more feasible for short sequences but is now becoming tractable for large proteins with adequate hardware resources. Furthermore, simulation runs are being extended into the microsecond time range with millisecond order runs predicted in the near future (Dror *et al.*, 2009; Klepeis *et al.*, 2009).

For the most part, *ab initio* simulations are not possible, however. Instead, starting structures for molecular mechanics-based simulations are generally generated via the method of homology modeling. The basis of this method is to generate a structure via an alignment with a homologue for which the structure is known. Where possible the co-ordinates of the template can be copied for identical portions of the alignment. For non-identical regions semi-empirical methods are used to generate starting co-ordinates. One of the most popular methods is by the satisfaction of spatial restraints, implemented by the MODELLER program. Probability distributions of protein structural features are derived from libraries of existing structures. From these, an objective function can be generated for model structure. Minimisation of the objective function thereby yields a structural model (Fig. 1.9). Closely related to homology modeling is threading (reverse-folding), which is based on the assumption that there are a limited number protein folds. By threading the model sequence through each structure and 3D database and scoring each possibility, remote homology can be detected and in some cases a reliable model structure can be inferred (Bourne and Weissig, 2003; Schlick, 2002).

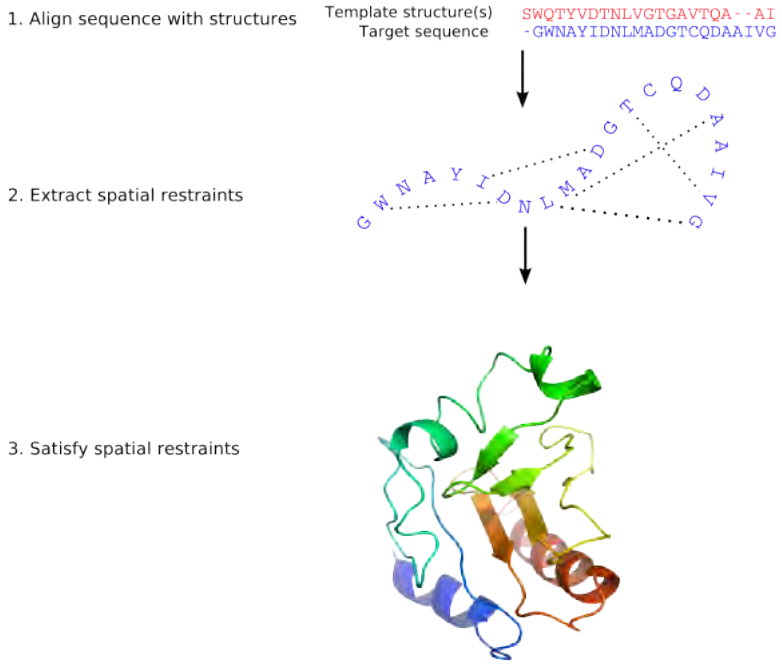


Figure 1.9: Homology modeling is implement in MODELLER.

1.3.3 *In silico* protein-protein docking

Many proteins function through their interaction with one another and other macromolecules such as RNA and DNA. It is predicted that protein-protein interactions are several orders of magnitude larger than the number of protein coding genes. A number of methods exist to predict protein-protein interactions. The yeast two-hybrid system allows for wholesale screening by using potentially interacting proteins to reconstitute a functional transcription factor linked to a reporter gene (Fields and Song, 1989; Stelzl *et al.*, 2005). Techniques designed for testing individual interactions such as affinity purification and chemical cross-linking can be combined with mass-spectrometry for large-scale mapping (Vasilescu and Figeys, 2006). Further experimental methods include tandem affinity purification, synthetic lethality, gene co-expression and protein arrays (Shoemaker and Panchenko, 2007). The number of protein complex structures remains extremely small, however. Wholesale structural determination of protein assemblies require combining a number of techniques (X-ray crystallography, NMR, cryo electron microscopy) and are still low throughput (Lensink *et al.*, 2007). Hence assembly structure determination is likely to lag behind individual structures for some time. The development of computational procedures to predict protein complexes from their individual components are therefore likely to play an important role in filling this gap. A number of non-structural methods exist to predict proteins are functionally and/or physically associated. These include gene neighbour and gene cluster detection, phylogenetic profiling, co-evolution, gene fusion (Rosetta Stone method), classification methods and Bayesian networks (Shoemaker and Panchenko, 2007). In this study the proteins in question are known to interact, the details of these interactions were therefore investigated using

structural methods. While computational methods have allowed homology modeling and in some cases *ab initio* structure prediction to become routine, protein-protein docking remains comparatively under-developed. Recent progress has been promising, however, as evidenced by the prediction competition CAPRI (Critical Assessment of PRedicted Interactions). During this competition blind prediction of a protein-protein complex is undertaken by various groups prior to the imminent release of the experimental structure (Méndez *et al.*, 2003, 2005).

A key problem in protein-protein docking is being able to successfully deal with protein flexibility and conformational change upon assembly formation. A successful program must be able to work with individual structures in the so-called unbound conformation to be of any usefulness. Most protein docking algorithms proceed via two stages: rigid docking of the individual proteins followed by further refinement of sidechains and/or protein backbone. Of the most popular methods used for the first stage is the representation of the protein surface on a cubic grid and the use of fast Fourier transforms or geometric hashing to determine geometric complementarity of the protein species (The general procedure is outlined in Fig. 1.10). During this phase all translations and rotations of the so-called mobile species are sampled and scored at predefined distance and angle intervals. Apart from geometric complementarity other features such as electrostatic interactions, Van der Waals interactions, hydrogen bonding and desolvation energies can be included to scoring. During the second stage, high scoring orientations of the first stage can be further refined using methods more akin to classical molecular mechanics based methods. This can include minimisation and molecular dynamics as well as optimisation of side-chain packing from rotamer libraries and backbone remodeling using Monte-Carlo sampling. Protein flexibility can also be handled by the rigid docking of multiple conformations of target protein, e.g. from molecular dynamics (Méndez *et al.*, 2003, 2005; Lensink *et al.*, 2007).

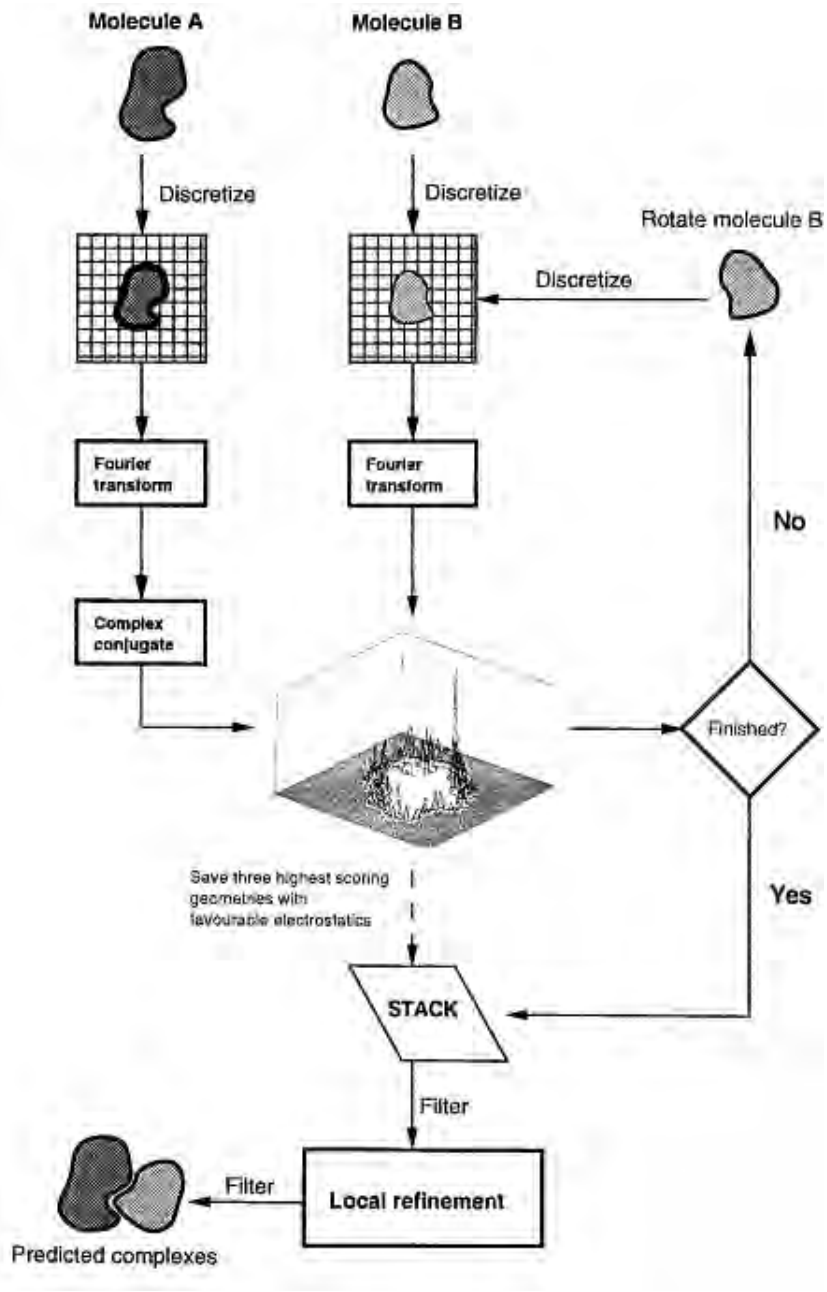


Figure 1.10: A standard protein-protein docking protocol. The molecules are each discretised on a 3D grid followed by the calculation of a correlation function using Fourier transformation. This is used to determine the degree of geometric and electrostatic overlap for a particular orientation. A 3D scan of all possible relative orientations is used to select high scoring positions for further refinement and filtering. Adapted from Gabb *et al.* (1997).

Protein-protein docking is still too unreliable to unambiguously predict the correct relative orientation and exact interface. However, despite the inability to predict the exact orientation, contacting interface residues are often correctly predicted. Furthermore, the various rounds of the CAPRI experiment have demonstrated that incorporation of pre-existing biochemical data often aids in providing accurate predictions. Protein docking predictions can therefore be used to direct course grained experiments such as site-directed mutagenesis

of potential interface residues. Predicted complexes can also be used to help solve low resolution structures, e.g. from cryo electron microscopy (Méndez *et al.*, 2003, 2005; Lensink *et al.*, 2007).

Increasingly protein-protein interactions are generating interest as therapeutic targets (Fuller *et al.*, 2009). Targeting protein-protein interactions present certain advantages compared to traditional active site based drugs. Firstly, targeting active sites may not always be the most appropriate approach. Secondly, an active site based drug may have detrimental effects due it's ability to also bind homologous proteins with similar functions. In the case of targeting pathogens, it is desirable to avoid binding to the host cognates for a drug to be effective. While often considered to be undruggable recent successes suggest this may not be the case. The licensed HIV-I drugs enfuvirtide and maraviroc that block viral entry demonstrate this (Melby and Westby, 2009). Proteins from the malaria parasite often possess unique features (discussed below) that make protein-protein interactions an attractive alternative target for novel drugs.

1.4 Malaria proteins as drug targets

1.4.1 Expression of malaria proteins

The rationalised identification of new inhibitors depends on possession of structural information. As for any other organism, the primary problem is obtaining high and pure protein yields for crystallisation trials. Recombinant expression of *Plasmodium* proteins in *E. coli* is notoriously difficult, however. A number of problems are typically encountered. The A+T richness results in substantially different codon usage compared to *E. coli*. *Plasmodium* genes are also typically much longer than their homologues in other organisms, as are the resulting proteins. Increased protein size is due mostly to long protein inserts with generally little homology to cognate enzymes. These inserts tend to be disordered and of low complexity, resulting in proteins that are not amenable to expression and crystallisation. Further problems include sporadic mutations of low complexity sequences introduced by *E. coli*, and cryptic prokaryotic translation start sites within *Plasmodium* genes. Improved levels of protein expression may be obtained by fine control of expression conditions such as a change of strain, addition of rare codon tRNAs or using a completely different expression system. Recently it has become more popular to express the target protein from synthetic genes coding for identical protein sequences but with a codon usage optimised for bacteria (Sugiyama *et al.*, 1996; Withers-Martinez *et al.*, 1999; Yadava and Ockenhouse, 2003; Flick *et al.*, 2004; Christopherson *et al.*, 2004). Mehlin *et al.* (2006) attempted a wholesale expression of 1000 *Plasmodium* genes and obtained soluble expression for only 63 genes. High predicted disorder, molecular weight, *pI* and lack of homology to *E. coli* proteins were all negatively correlated with soluble expression.

1.4.2 Existing structures

The difficulty of expressing *Plasmodium* proteins is reflected by the paucity of structures in the Protein Data Bank (Kihara and Skolnick, 2003). As of June 2009, querying the PDB (http://www.pdb.org) for structures of *Plasmodium* proteins and excluding sequences with greater than 90% identity, yields 118 entries (de Beer *et al.*, 2009). A closer inspection of all released *Plasmodium* protein structures reveals 100 orthologues from multiple *Plasmodium* species. In contrast, querying the PDB for human protein entries (excluding $> 90\%$ sequence identity) reveals more than 4500 structures. Even though the number of *Plasmodium* protein structures is still alarmingly sparse, there has been an almost doubling in *Plasmodium* protein structures since 2005, largely due to the advent of structural genomics programs including the Structural Genomics Consortium, (http://sgc.utoronto.ca) and the Structural Genomics of Pathogenic Protozoa (http:// www.sgpp.org). The Structural Genomics Consortium (SGC) reported 25 distinct *Plasmodium* protein crystal structures from five species. The success rate of this study is similar to other structural genomics programs, and demonstrates the viability of structural genomics for protozoa. This was partly due to treating orthologues from multiple species as alternative expression constructs (Vedadi *et al.*, 2007). The SGPP Consortium has solved 40 structures from the parasitic organisms Leishmania, *Trypanosoma brucei*, *T. cruzi* and *Plasmodium* of which 16 are *Plasmodium* proteins. The success is attributed to pioneering a number of developments such as domain prediction, the use of co-crystallents, capillary crystallisation and "fragment cocktail crystallography".

1.4.3 Modeling of *Plasmodium* proteins

In lieu of the paucity of crystal structures for *Plasmodium* proteins it is often necessary to resort to homology modeling. This approach depends critically on the alignment with template structures. Unfortunately the biased nucleotide and amino acid composition (Bastien *et al.*, 2004b) and *Plasmodium*-specific inserts make it difficult to correctly identify core-conserved regions. The presence of inserts often confuses multiple and structural-alignment programmes. A number of techniques have been used to circumvent this problem (Fig. 1.11). From a first pass alignment, approximate insert positions can be determined. Sequences can then be split according to long inserts and re-aligned. Inserts can vary considerably across different *Plasmodium* species (Birkholtz *et al.* 2004 and C. Claudel-Renard, personal communication). While adjusting an alignment for modeling, it is useful to refer to phylogenetically diverse multiple alignments including as many *Plasmodium* sequences as possible (Wells *et al.*, 2006). As an adjunct to alignment, independent motif discovery (e.g. with MEME, Bailey and Elkan 1994) can be used to fix mistakes that alignment programmes frequently make when aligning long *Plasmodium* proteins with homologues (Wells *et al.*, 2006; de Beer *et al.*, 2006). Further improvements can be made by using hydrophobic cluster analysis (Callebaut *et al.*, 2005) and secondary structure predictions to align homologous regions within inserts. Once an alignment has been decided on, based on visual assessment,

a series of models can be built. Because of the high degree of uncertainty that often accompanies alignments used for modeling *Plasmodium* proteins, it is usually not feasible to rectify all structural anomalies. But by performing standard quality checks on a large sample of models and summarising the results, it is possible to identify parts of the alignment causing most problems. Refined alignments might benefit from species-specific matrices that take into account the differences of amino acid distribution between the aligned proteins (Bastien *et al.*, 2004a, 2005).

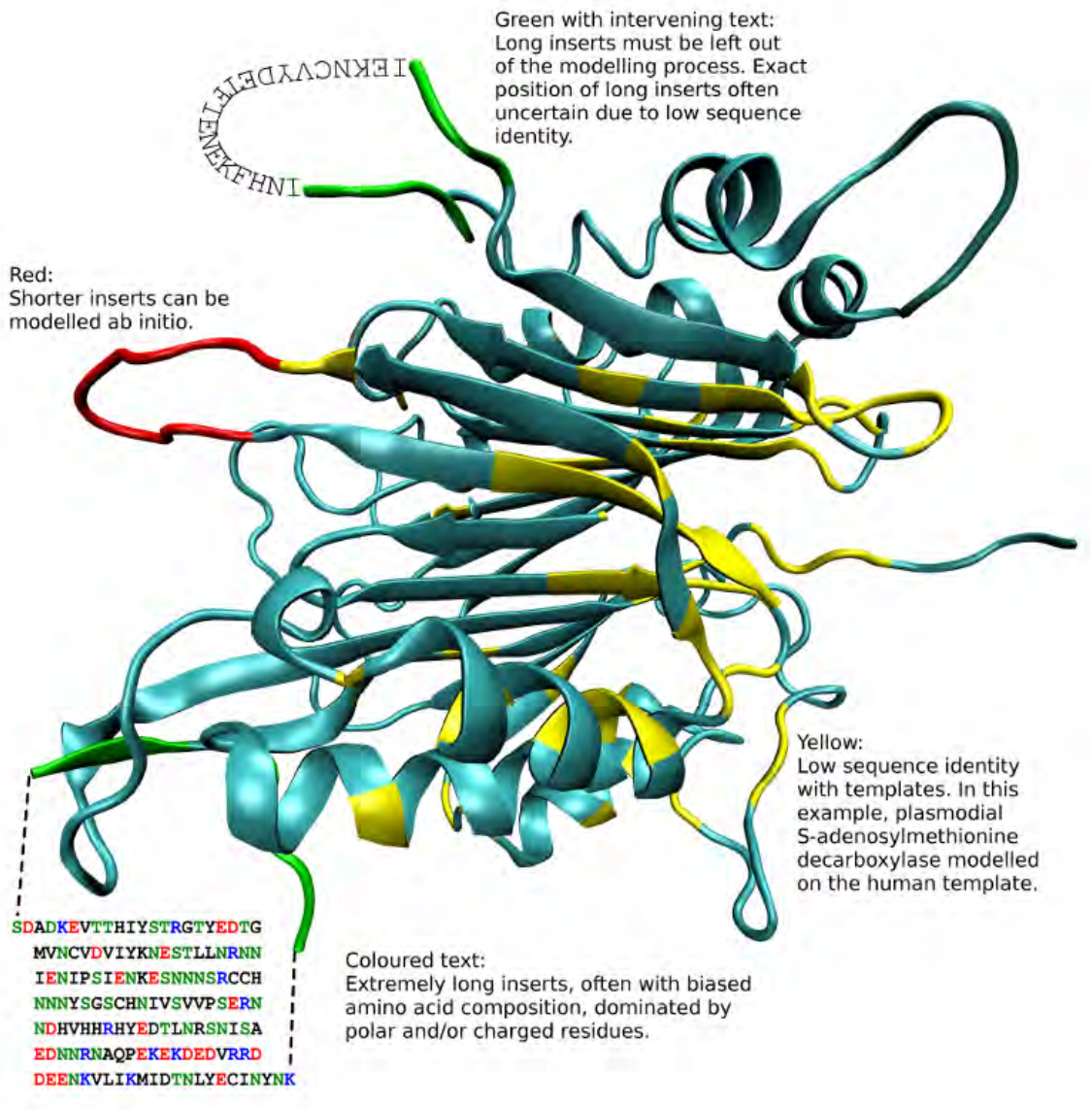


Figure 1.11: Problems frequently encountered with modeling of *Plasmodium* proteins. See text for further details.

Despite the difficulties with homology modeling of *Plasmodium* proteins there have been some notable successes. *P. falciparum* DHFR forms part of a bifunctional protein that also

carries thymidylate synthase. A number of existing drugs such as cycloguanil and pyrimethamine target the DHFR domain, and have been used effectively in the past. However, drug resistance has evolved that reduces the usefulness of this important class of drugs. Hence *P. falciparum* DHFR has been a popular target for homology modeling efforts (Toyoda *et al.*, 1997; McKie *et al.*, 1998; Lemcke *et al.*, 1999; Rastelli *et al.*, 2000; Santos-Filho *et al.*, 2001; Delfino *et al.*, 2002). Toyoda *et al.* (1997) were able to identify new inhibitors in the micromolar range. McKie *et al.* (1998) and Lemcke *et al.* (1999) could rationalise the pyrimethamine resistance caused by the Ser 108 Asn mutation. One of these models was further used to identify new inhibitors acting in the nano- and micromolar ranges (McKie *et al.*, 1998). Delfino *et al.* (2002) in turn used their model to investigate a large number of antifolate resistant mutants. Rastelli *et al.* (2000) further explained the cycloguanil resistance/pyrimethamine sensitivity conferred by Ala 16 Val/Ser 108 Thr, as well as the ability for WR99210 to inhibit both pyrimethamine and cycloguanil resistant mutants. A number of new inhibitors were also successfully designed. The high accuracy of the alignment used for modeling meant that predicted dockings were subsequently confirmed with the crystal structure of the complete bifunctional enzyme (Yuvaniyama *et al.*, 2003).

Considerable work has also gone into modeling *Plasmodium* proteases essential to the parasite's intra-erythrocytic life stage. A number of these models have been used to identify new inhibitors (Li *et al.*, 1996; Desai *et al.*, 2004, 2006; de Terán *et al.*, 2006b), although the increasing number of crystal structures for these proteases is likely to gradually replace the need for homology models. de Terán *et al.* (2006a) demonstrated the advantages of using multiple structures with plasmepsin IV from *P. falciparum*. A homology model and a low resolution crystal structure were both used for inhibitor identification. The homology model performed better on structural quality indicators and was more robust when calculating binding energy for an inhibitor series. The enhanced structural quality of the homology model was put down to the intermediate resolution of the X-ray structure (2.8 Å). Further improvements in predicting binding were gained by using a combined model employing both structures, as well as using molecular dynamics to increase sampling. The improved docking performance argues for making use of multiple experimental and predicted models instead of relying on a single structure (Luksch *et al.*, 2008).

Singh *et al.* (2004) used homology modeling to derive a chimeric berghepain-2 that more closely resembled falcipain-2 in it's sensitivity to inhibitors. The motivation behind this approach was to create an *in vivo* rodent model of the *P. berghei* protein that mimics this important human drug target in *P. falciparum*. Homology modeling with molecular dynamics was used to predict the structure, substrate binding and MOA of histo-aspartic protease from *P. falciparum* (Bjelic and Aqvist, 2004). Other noteworthy examples include homology models of dihydropteroate synthase (DHPS) from *P. vivax* and *P. falciparum* to explain the refractory nature of the *P. vivax* enzyme to sulfadoxine (Korsinczky *et al.*, 2004). A homology model of histone deacetylase 1 from *P. falciparum* was successfully used to identify inhibitors in the nanomolar range with significant selectivity compared to mammalian cells (Andrews

et al., 2008). Homology models combined with molecular dynamics were used to explain sulfadoxine resistance in mutants of *P. falciparum* DHPS (Rastelli *et al.*, 2000).

A remarkable achievement is exemplified by the homology model obtained for *P. falciparum* farnesyltransferase (Ras FTase) based on a rat homologue (Glenn *et al.*, 2005). The sequence identity between the target and template was quite low (23%) including a parasite-specific insert of approximately 100 residues in the *Plasmodium* protein. Using this model in the docking program GOLD, a range of ethylenediamine based inhibitors with $IC_{50} < 50\ nM$ were identified of which two had an IC_{50} of less than 1 nM. This range of inhibitors was subsequently used together with the model for further rounds of optimisation to derive new structures with better selectivity (up to 145 fold) towards the *P. falciparum* enzyme compared to its mammalian counterpart. Preliminary pharmacokinetics promisingly indicated that some of the compounds were metabolically stable (Glenn *et al.*, 2005, 2006; Fletcher *et al.*, 2008). The results of this work are encouraging and demonstrate that low sequence identity and the presence of inserts need not be a barrier to inhibitor discovery.

1.5 Summary and aims

Due to it's prevalence and increasing drug resistance malaria remains a pressing world health problem that requires urgent attention. Due to increasing resistance, the identification of new drugs remains urgent. This will be best facilitated by a greater understanding of the parasite's basic biology. The high attrition rate of potential drug leads late in the research phase has created the need for rational drug design. This approach will benefit most on gaining structural knowledge of the parasite's macromolecules that represent promising targets for inhibition. However, *Plasmodium* proteins possess a number of characteristics that make structural determination difficult using current experimental methods. To fill the gap computational methods can be applied to facilitate further experiments designed to understand and possibly exploit these unique characteristics. The enzymes of the *P. falciparum* polyamine pathway have been identified as a potential drug target and also exemplify this unique characteristic of *Plasmodium* proteins. Compared to the human host, the arginase of *P. falciparum* displays a strong dependency between trimerisation, enzyme activity and the presence of the active site metals. Additionally, the bifunctional arrangement of *S*-adenosylmethionine decarboxylase/ornithine decarboxylase is apparently unique to *Plasmodium* and is definitely absent in the human host. This study describes the application of various molecular modeling techniques to further understand the unique characteristics of these enzymes. Homology modeling and molecular dynamics of arginase revealed a novel inter-monomer interaction that is involved in the structural metal dependency. This interaction may serve as a potential parasite-specific target. Homology modeling and docking of AdoMetDC and ODC from five *Plasmodium* species was pursued to predict the quaternary structure of the bifunctional complex. Conserved regions and specific residues were identified as likely candidates for mediating AdoMetDC/ODC binding, and have targeted for further experimental follow-up.

The findings discussed can and have been used to guide further experimental analysis that may ultimately lead to novel therapeutic exploitation of these proteins.

Partof this work has been published in the FEBS Journal (Wells *et al.*, 2009) and presented at the following conferences:

- Investigations into the structural metal dependency of malarial arginase with molecular dynamics. Intelligent Systems for Molecular Biology (ISMB). July 2007, Vienna, Austria. 2.

- Investigations into the structural metal dependency of malarial arginase with molecular dynamics. First African Structural Biology Conference. November 2006, Wilderness, South Africa.