Chapter 3. DfrB Dihydrofolate Reductase

Introduction and wide use of novel synthetic drugs for the control and treatment of bacterial infections has resulted in the emergence of resistant organisms that have developed novel catalytic activities, either modifying old enzymes or recruiting new ones, to meet new environmental challenges and survive under selective pressure [468]. DfrB, or type II dihydrofolate reductase (DHFR), is an example of such an enzyme. It is almost completely insensitive to the widely used antifolate drug trimethoprim (TMP) [469], which inhibits bacterial DHFR effectively.

Understanding of the approach DfrB DHFR uses to bind the ligands and promote the reaction, as well as the origins and general properties of the DfrB family, is not only of clinical relevance but also important as a basic research question. On the one hand, the presence of this enzyme within integrons located in mobile elements [470-473] facilitates the propagation of antibiotic resistance through horizontal transfer and, on the other hand, despite catalysing the same reaction as the chromosomal DHFR it has neither sequence nor structural homology with it.

The following sections detail the work I have carried out on DfrB DHFR in order to provide a deeper understanding of this family of proteins, its origins and catalytic properties. First, I combined sequence and structural comparisons with a detailed overview of recent functional studies; this analysis highlights the possible origins of the DfrB family and its atypical properties. In this first section I also introduce the concept of integrons and gene cassettes, as well as the general properties of the SH3 folding domain. This is followed by a comprehensive computational study on the binding modes of the ligands dihydrofolate (DHF) and NADPH within the active site, and their dynamic behaviour.
3.1 Resistance to trimethoprim

Tetrahydrofolate (THF) and its derivatives are essential cofactors involved in the metabolism of serine, glycine, and methionine, among the amino acids, and also used in the synthesis of purine and thymine nucleotides. Because of its metabolic importance, tetrahydrofolate synthesis pathways constitute a primary target for antimicrobial, antiparasitic and anticancer (i.e. cytotoxic) chemotherapy.

The enzyme dihydrofolate reductase (DHFR) catalyses the reduction of dihydrofolate (DHF) to tetrahydrofolate (THF) using nicotinamide adenosine diphosphate (NADPH) as cofactor. As it is one of the most conserved enzymes among living organisms [474], it constitutes an interesting target for broad-spectrum drug design. TMP is a synthetic drug introduced for clinical use in western Europe in the early 60s [475]. Whereas its structural similarity to DHF makes it a competitive inhibitor of the ubiquitous chromosomal DHFR in bacteria, fungi and protozoa, mammalian DHFR is resistant to TMP. Therefore, the specificity and selectivity of this antifolate drug [476] have lead to its widespread use in the treatment of human infections.

Soon after the clinical introduction of TMP the first cases of resistant bacteria began to emerge, including both chromosomal and plasmid-borne resistance [477]. The most common mechanism of resistance involves an alternative DHFR enzyme encoded within mobile elements (e.g. plasmids and transposons), which are rapidly spread within a bacterial community by horizontal transfer [477]. The first plasmid-mediated resistance to TMP was reported in 1972 [478], and since then an increasing number of plasmid-encoded DHFR genes (dfr) have been found. They are grouped into two main families, A and B [479].

The dfrA gene family is diverse and encodes proteins of 152 to 189 amino acids, with identity levels between 20 to 90%, and some structural and sequence similarities to the chromosomal enzyme. There are at least 20 different dfrA sequences reported so far ([480] and references therein). The dfrB gene family, on the other hand, encodes a unique group of enzymes, referred to as DfrB, which in terms of sequence and structure are completely unrelated to the chromosomal or other DHFRs. The dfrB genes code for similar and much shorter proteins (78 residues) with identity levels of 75% or above, which are
extremely resistant to TMP ($K_{inhibition} \sim 7500$ larger than that of the chromosomal DHFR) \cite{469}. Six different $dfrB$ genes, isolated from several organisms in various countries (see Table 3-1), have been reported and deposited in the GenBank database \cite{481} in the last 30 years.

### Table 3-1 Collection of $dfrB$ genes reported to the GenBank.

<table>
<thead>
<tr>
<th>$dfrB$ gene</th>
<th>Year of isolation</th>
<th>Country</th>
<th>Host organism</th>
<th>GenBank accession number</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$dfrB1$</td>
<td>1972-1974</td>
<td>France</td>
<td><em>Escherichia coli</em></td>
<td>K02118</td>
<td>[482]</td>
</tr>
<tr>
<td></td>
<td>1995</td>
<td>Sweden</td>
<td><em>Escherichia coli</em></td>
<td>U36276</td>
<td>Unpub.*</td>
</tr>
<tr>
<td></td>
<td>2001</td>
<td>Germany</td>
<td>Uncultured bacterium, environmental sample</td>
<td>AY139601</td>
<td>[483]</td>
</tr>
<tr>
<td></td>
<td>2000-2003</td>
<td>Germany</td>
<td><em>Bordetella bronchiseptica</em></td>
<td>AJ879564</td>
<td>[484]</td>
</tr>
<tr>
<td></td>
<td>2003</td>
<td>Spain</td>
<td><em>Escherichia coli</em></td>
<td>AY970968</td>
<td>[485]</td>
</tr>
<tr>
<td>$dfrB2$</td>
<td>1972</td>
<td>UK</td>
<td><em>Escherichia coli</em></td>
<td>U12441</td>
<td>[486]</td>
</tr>
<tr>
<td></td>
<td>2001</td>
<td>Germany</td>
<td>Uncultured bacterium, environmental sample</td>
<td>AY139592</td>
<td>[483]</td>
</tr>
<tr>
<td>$dfrB3$</td>
<td>1972-1973</td>
<td>UK</td>
<td><em>Klebsiella aerogenes</em></td>
<td>X04128</td>
<td>[487]</td>
</tr>
<tr>
<td></td>
<td>1985</td>
<td>Australia</td>
<td><em>Escherichia coli</em></td>
<td>AY123252</td>
<td>[488]</td>
</tr>
<tr>
<td></td>
<td>1988-1990</td>
<td>Scotland</td>
<td><em>Aeromonas salmonicida</em></td>
<td>AF327729</td>
<td>[473]</td>
</tr>
<tr>
<td>$dfrB4$</td>
<td>2001</td>
<td>Sweden</td>
<td><em>Escherichia coli</em></td>
<td>AJ429132</td>
<td>[489]</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>China</td>
<td><em>Aeromonas hydrophila</em></td>
<td>AY751518</td>
<td>Unpub.</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>China</td>
<td><em>Klebsiella pneumoniae</em></td>
<td>AY968808</td>
<td>Unpub.</td>
</tr>
<tr>
<td></td>
<td>2004</td>
<td>China</td>
<td><em>Escherichia coli</em></td>
<td>AY973253</td>
<td>Unpub.</td>
</tr>
<tr>
<td>$dfrB5$</td>
<td>2004</td>
<td>USA</td>
<td><em>Pseudomonas aeruginosa</em></td>
<td>AY943084</td>
<td>[490]</td>
</tr>
<tr>
<td>$dfrB6$</td>
<td>2000-2001</td>
<td>Australia</td>
<td><em>Salmonella enterica</em></td>
<td>DQ274503</td>
<td>[491]</td>
</tr>
</tbody>
</table>

* Unpublished

The latest reported sequence corresponds to a new class of $dfrB$ gene, $dfrB6$, which was found in a $dfrB6$-aadA1 cassette array in a class 1 integron, isolated from several multiply antibiotic-resistant *Salmonella enterica* serovar Infantis strains in Australia \cite{491}.

A simple decrease in the usage of TMP has been ineffective in controlling the problem of rising resistance \cite{492}. A better understanding of the underlying mechanism of resistance acquisition and the enzymes involved seems critical to prevent or delay the emergence of other drug-resistant pathogens. The origin of the $dfrB$ genes is still a mystery. They have been found not only in human pathogens, but also in bacteria that affect fish \cite{473}, pigs \cite{484,493}, cows \cite{494}, and also in wastewater samples \cite{483}. The worldwide spread of these enzymes
and their occurrence in various bacterial genera arises from their presence in broad-host mobile elements; \textit{dfrB} genes have been found only within gene cassettes in integrons [479].

### 3.2 R67 vs chromosomal DHFR

There is neither sequence nor structural homology between DfrB and its chromosomal counterpart, despite both enzymes catalysing the reduction of dihydrofolate (DHF) to tetrahydrofolate using NADPH as cofactor (Table 3-2) (for a recent review comparing both enzymes see [469]).

<table>
<thead>
<tr>
<th>Property</th>
<th>\textit{E. coli} DHFR</th>
<th>DfrB DHFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>\textit{Escherichia coli}</td>
<td>Trimethoprim resistant bacteria. Integron - gene-cassette encoded [471,495,496]</td>
</tr>
<tr>
<td>Major physiological role</td>
<td>Reduces folate and DHF to THF [100,497,498]</td>
<td>Reduces folate and DHF to THF. Confers resistance to the antibiotic trimethoprim [471,496,499]</td>
</tr>
<tr>
<td>Substrates</td>
<td>Folate ((k_{cat}= 0.0037 \text{ s}^{-1})) [500])</td>
<td>Folate ((k_{cat}= 0.0036 \text{ min}^{-1}) [502])</td>
</tr>
<tr>
<td></td>
<td>DHF ((k_{cat}= 240 \text{ s}^{-1})) [501])</td>
<td>DHF ((k_{cat}= 1.3 \text{ s}^{-1}), [503])</td>
</tr>
<tr>
<td>Cofactor</td>
<td>NADPH</td>
<td>NADPH</td>
</tr>
<tr>
<td>Structure</td>
<td>Monomer (see Table 1 in reference [100])</td>
<td>Dimer of dimers, toroidal structure with a central active pore [504,505]</td>
</tr>
<tr>
<td>Active site polarity</td>
<td>Non polar [506]</td>
<td>Positively charged [507]</td>
</tr>
<tr>
<td>Binding/release sequence</td>
<td>Ordered ternary complex mechanism [100,497,498]</td>
<td>Random [502]</td>
</tr>
<tr>
<td>Reduction mechanism order (proposed)</td>
<td>1(^{\text{st}}) protonation 2(^{\text{nd}}) hydride transfer [508,509]</td>
<td>1(^{\text{st}}) protonation 2(^{\text{nd}}) hydride transfer [510]</td>
</tr>
<tr>
<td>Stereochemistry of the reaction</td>
<td>\textit{pro-R} hydrogen, \textit{endo} conformation of the rings [511]</td>
<td>\textit{pro-R} hydrogen, \textit{exo} conformation of the rings [507,512,513]</td>
</tr>
<tr>
<td>Rate-determining step</td>
<td>Product release [497]</td>
<td>Hydride transfer [503]</td>
</tr>
</tbody>
</table>

Whereas the chromosomal enzyme has several features that facilitate the reaction, including a complex well-conserved binding site for each ligand and the existence of electrostatic and conformational components that facilitate the reaction, these seem to be absent in the simpler DfrB enzyme. DfrB is enzymatically 1000 times less efficient than chromosomal DHFR [469].
Figure 3-1 Crystal structure of *E. coli* DHFR (PDB code 1RA2) (A) and R67 DHFR (PDB code 1VIF) (B).

The protein structure is represented using secondary-structure cartoon, while the ligands bound within the active site are depicted as stick models. While the complete structures of folate and NADPH were resolved in the structure of the monomeric *E. coli* DHFR [100], only the pterin ring of folate is visible in the tetrameric structure of R67 DHFR [505].
The chromosomal-encoded DHFR, or type 1 DHFR, is a biologically ubiquitous enzyme that has been extensively studied since the 1950s [514]. There have been major advances in elucidating its kinetic pathways, catalytic mechanism and structure [100,497]. On the other hand, very little is known about the DfrB DHFR, a type 2 enzyme isolated only in the mid 1970s [471]. This kind of DHFR has been found only as gene cassettes in integrons [515] and does not have any known homologues (see below sections 3.4 and 3.5). Not only are its sequence and 3-D structure completely unrelated to those of the conventional DHFR enzyme (Figure 3-1), but also they seem to present a substantially different catalytic mechanism.

While the type 1 DHFR catalyses the transfer of the pro-R hydride ion from the nicotinamide ring of NADPH to the pterin ring of DHF in an endo conformation [103], the DfrB enzyme promotes an exo reaction [512] (Figure 3-2). This implies that the relative orientation of the ligands inside the active site and, therefore, the potential energy surface for the reaction, must be significantly different between the two enzymes.

![Figure 3-2 Stereochemistry of the reaction.](image)

Spatial disposition of the reactive rings of pterin, from DHF (grey), and nicotinamide, from NADPH (black), within the chromosomal DHFR (A) and DfrB (B) enzyme. While in both cases it is the pro-R hydrogen that is transferred (shown in red), the rings adopt an endo conformation within the active site of the type 1 DHFR, and an exo conformation in the DfrB pore.
Moreover, while the active site of the monomeric chromosomal enzyme is a narrow binding groove where both reactants DHF and NADPH bind to different specific locations [516], DfrB DHFR presents a spacious active-site cavity formed by four identical subunits, without a distinctive binding position for each ligand. Furthermore, as the DfrB enzyme does not seem to possess a proton donor group in the active site [505], it has been suggested that it may directly use protonated DHF (DHFH\(^+\)) as substrate [510].

### 3.3 Motivations and Goals

The DfrB family of proteins constitute a unique group of enzymes. They are involved in the widespread and fast distribution of trimethoprim resistance, and very little is known in regards to their evolutionary origin and mechanism of action.

The first question I tried to address was the binding of the reactants dihydrofolate (DHF) and NADPH to the active site of R67 DHFR, a DfrB enzyme. The structural characterization of a reactive ternary complex R67•DHF•NADPH had proven to be elusive to experimental analysis so far and, therefore, a series of computational approaches were employed to provide further insight into this system. A combination of docking studies and molecular dynamics simulations were applied, leading to a reasonable structure for the reactive ternary complex and providing significant understanding of the way in which R67 binds and interacts with its ligands [243].

The peculiar structure of R67 DHFR, combined with the flexible binding of the ligands, clearly suggested that this was an unusual enzyme. To further understand the system, a series of sequence and structural comparisons were carried out. Database searches, and detailed analysis and compilation of published data, produced an integrated body of work that allowed me to put the DfrB family of enzymes into an evolutionary context. Their exclusive presence within integrons, as gene cassettes, their homotetrameric structure composed of SH3 domains, and their tolerance to mutations, strongly supports the idea of these being a non-conventional group of enzymes. A group of proteins that have probably been evolving for a long period of time, and came to light only recently as a novel response to increasing antibiotic pressure. A selected part of
this work has been recently published [517], and a complete description of all my studies is presented below.

In the following sections I present a detailed account of my work, not in the chronological order it was done but following a more rational scheme. To facilitate the understanding of the problem, I introduce first the sequence analysis of the DfrB DHFR family of enzymes and their genetic framework, integrons and gene-cassettes, followed by structural studies and a description of the SH3 domain. Finally, I describe my initial studies on ligand binding and protein-ligand interactions, and the generation of a model of the reactive ternary complex.

3.4 Genetic framework and sequence analysis

In order to shed light on the history and evolution of the DfrB family of proteins, a series of sequence and structural analyses were carried out. Database searches were used to identify all reported sequences; these were then compiled and analysed. Their wide distribution over several bacterial genera and their rapid spread clearly correlates with their unique presence in integrons, in which they are encoded as gene cassettes. Structural homologues were found in public databases, and the characteristics of the SH3 domain provided further support for the proposed relative crudeness of this catalyst.

3.4.1 Methods

Sequence analysis of the DfrB family of proteins began with the identification of all members of the family reported to GenBank [481]. The protein sequence of R67 DHFR, a DfrB protein, was used as the primary template. The NCBI BLAST facility [518] was used to perform a tblastn (protein query vs. translated database) of the complete database. All hits were manually inspected and the different sequences grouped into family members. Six different members of the DfrB family were identified (Figure 3-4), all of them encoded as gene cassettes within integrons. In order to find possible ancestors of the DfrB proteins a search was also performed in the TIGR database [519]; there were no successful findings. Similarly, sequences homologous to the attC recombination site of the
DfrB dihydrofolate reductase

dfrB cassettes were not found either in GenBank or TIGR, limiting the possibilities of finding the possible host ancestor of this family.

The program pfafat (Protein Family Alignment Annotation Tool) [520] was used to analyse and annotate the complete nucleotide sequence of all six different DfrB cassettes identify. MEGA 3.1 [521] was used to construct a phylogenetic tree of the DfrB family. The minimum evolution algorithm, with all default parameters, and a 500 replicates bootstrap analysis [522] was performed to generate a final consensus phylogenetic tree.

The Nei-Gojobori method, as implemented in the program MEGA 3.1 [521], was used to compare the synonymous (silent, \(d_S\)) and nonsynonymous (amino acid-changing, \(d_N\)) substitution rates for each pair of dfrB sequences. The number of synonymous (\(S_d\)) and nonsynonymous (\(N_d\)) sites are counted, and normalized according to the total number of possible synonymous (\(S\)) or nonsynonymous (\(N\)) sites. Finally, a correction for multiple substitutions at the same site is made (\(d_S\) and \(d_N\)).

\[
d_S = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p_S \right) \quad d_N = -\frac{3}{4} \ln \left(1 - \frac{4}{3} p_N \right)
\]

\[
p_S = S_d / S \quad p_N = N_d / N
\]

A positive selection is expected to produce \(d_N > d_S\), while a purifying selection is reflected by \(d_N < d_S\).

The online DNA mfold server [523] was used to predict the putative secondary structure of the attC recombination site.

**3.4.2 Genetic Framework: Integrons and Gene Cassettes**

All known dfrBs are cassette-encoded. Gene cassettes are small mobile elements that generally include only a single open reading frame (ORF) and a specific recombination site, attC [524]. They can exist either as free circular molecules, usually transcriptionally inactive, or inserted within integrons as linear molecules [525] (Figure 3-3).
Integrons constitute natural cloning systems that contain a recombination site, *attI*, where different gene cassettes can be inserted by a site-specific integrase encoded within the integron, *intI* [526,527] (Figure 3-3). Once inserted, most cassettes can become transcriptionally functional, relying on a promoter region within the integron for expression. The position of the cassette within the inserted region has a strong effect on cassette expression, as those cassettes closer to the promoter will be expressed at higher levels than those further from it. Cassettes do not have an origin of replication and, therefore, can replicate only when inserted within an integron, which, in turn, relies on its genetic framework, usually a plasmid or chromosome, for its own replication.

**Figure 3-3 Schematic representation of the structure of an integron and mobile gene cassettes.**

Gene cassettes usually consist of two functional components, an open reading frame (ORF) and an *attC* recombination site. They can exist either as a free circular non-replicative element, or as an active gene within an integron. An integrase (IntI) is involved in the recruitment and mobilization of gene cassettes; it recognizes the *attC* site and catalyses the integration of the cassette at a specific site (*attI*). A promoter region (Pc) located adjacent to the *attI* recombination site within the integron is used for the expression of the inserted cassettes. Thus, cassettes closer to the promoter region have higher levels of expression than those more distal in the array to the *attI* site. Arrows below genes indicate direction of transcription.

More than 70 different cassettes conferring resistance to a broad range of antibiotics have been found (see [528] and references therein). The immediate
origin of the resistance genes encoded within cassettes is not clear. From an evolutionary viewpoint, the emergence of antibiotic resistance within a short period of the introduction of antibiotics suggests these ORFs were acquired from a pre-existing genetic reservoir. This is supported by findings of antibiotic-resistance genes in uncultured bacteria from natural habitats [529,530]. More than 10,000 different natural antibiotics produced by actinomycetes alone have been catalogued [531], any number of which may have been the original substrates for the proteins coded in gene cassettes [528]. Resistance-gene homologues have been found in the cassettes of a *V. metschnikovii* strain isolated in 1880 [532], 40 years before the discovery of penicillin.

Most integrons characterized to date are found in mobile elements such as plasmids and transposons. They are divided into class 1, 2 or 3 based on the encoded integrase, with class 1 being the most common among multidrug-resistant bacteria [528,533]. It is thought, however, that they were originally present within bacterial chromosomes and became embedded in mobile elements at a later stage [534]. Although the precise ancestors have yet to be identified, some chromosomal integrons have been reported [535-537]. Phylogenetic analysis reveals that the extent of divergence between the *intI* genes largely adheres to the line of descent among the bacterial species, suggesting that chromosomal integrons are ancient genetic structures that can harbour up to several hundred cassettes, most of which encode for proteins with no known homologues or function.

The noticeable variability in codon use and base composition in the ORF region of these cassettes suggests that they have a heterogeneous origin, including from bacteria, viruses and eukaryotes. The *attC* recombination sites, however, seem to be closely related among cassette arrays within a specific chromosomal integron [528,534], suggesting that cassettes are assembled within particular bacteria using a still unknown mechanism. Their special characteristics, such as the lack of a promoter region and the presence of the *attC* recombination site suggest that there is a specific mechanism for their construction. It has been proposed that they may originate from mRNA, through a reverse transcriptional process, and that the *attC* element may be added latter [538]. However, this mechanism does
not provide a plausible explanation for the generation of those few cassettes that do contain a promoter region, or are inserted in the opposite direction [528,539].

Gene cassettes have the potential to be recruited by class 1, 2 and 3 integrons. As integrons within mobile elements move from host to host through horizontal transfer, not only do they spread antibiotic resistance but they could become exposed to new cassettes, which might be incorporated into the mobile genetic repertoire [540], as it has been shown that some integrases can catalyse both the insertion and removal of gene cassettes from different integrons [540]. Although this gene exchange system has been mainly associated with the emergence and fast distribution of antibiotic resistance, it is clear that its function within bacterial communities has yet to be defined and it may be far more complex, and could have an important role in bacterial evolution and survival [541,542].

There are several cases in which closely related genes are associated with closely related attC recombination sites, indicating that they may have evolved from a common ancestor. However, the degree of divergence in the coding regions of these cassettes is within the range found for pairs of equivalent genes from *E. coli* and *Salmonella typhimurium* [543], which are classified as separate species and are estimated to have diverged between 10 and 160 millions years ago [544]. Thus, if the pairs of cassettes are derived from common ancestors, cassettes are likely to be very ancient structures and the attC sites must have been functionally conserved over very long periods.

### 3.4.3 Sequence analysis: the DfrB Cassettes

I found six different DfrB enzymes in total during database searches (Figure 3-4). Their level of sequence identity ranges from 74% (DfrB2, DfrB3 and DfrB4) to 88% (DfrB1 and DfrB5). Most of the variability is located in the N-terminal region (Figure 3-9A), which has been shown not to be essential for enzyme catalysis [503,505]. Residues that have been suggested to be important for protein catalysis (Lys32, Gln67, Ile68 and Tyr69 [505,545-548]) appear to be conserved among the DfrB family.
Chapter 3  
DfrB dihydrofolate reductase

<table>
<thead>
<tr>
<th>1</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DfrB1</strong></td>
<td>MERSINEVSNP</td>
<td>VAGNPVFPSATPS</td>
<td>MDVRK</td>
<td>KSQA</td>
<td>MQ</td>
<td>Q1Y</td>
<td>PFVAALERIN</td>
</tr>
<tr>
<td><strong>DfrB3</strong></td>
<td>-DGHN.G..TL...Q.A.L...H...L...V...K...S...VA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DfrB4</strong></td>
<td>-NEGO...TSA..R.A..AL...R...T...S...MT...VA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DfrB5</strong></td>
<td>-DQGRS...Q.A...A...K...S...VA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DfrB6</strong></td>
<td>-DQG...I...Q.AS...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3-4 Sequence alignment of the six known DfrB proteins.

DfrB1 has been used as the reference protein (see Table 3-1 for details of sequences); only mutated residues are shown for the other sequences, with conserved residues represented by dots. Residues reported to be associated with ligand binding and catalysis are shown in bold. It is clear that most mutations cluster at the N-terminal end, which has been shown not to be required for catalytic activity.

Comparisons of homologous proteins from *Escherichia coli* and *Salmonella enterica* indicate an average rate of sequence divergence of ~1% per million years [549]. Although the mutational rate of non-essential proteins is expected to be higher, the ~20% sequence divergence between some of the *dfrB* genes isolated after just 10 years of TMP use suggests that they already existed well before the introduction of the drug, and had long been diverging before coming to light as a result of drug pressure.

Under stress conditions, such as the presence of an antibiotic, the level of mutagenesis of those genes that are being extensively transcribed or that have a direct effect on the relief of the stress is expected to increase [550-552]. While in normal conditions most of the mutations are likely to result in synonymous substitutions (e.g. no net effect on the final protein sequence), stress conditions will elevate the number of nonsynonymous mutations (e.g. producing net changes on the final protein) [553]. It has been observed that antibiotics not only select for resistant strains but also act as indirect promoters of antibiotic resistance by increasing the mutation frequency [554,555]. In the particular case of TMP, it has been shown to cause a 10-fold increase in mutation frequency, mainly as base substitutions and deletions in wild type *E. coli* [552]. It has also
been shown recently that DHFR inhibitors accelerate the evolutionary process of chromosomal DHFR of *Pneumocystis jirovecii* by increasing the number of nonsynonymous mutations over synonymous mutations (positive selection) [556].

Therefore, to determine if the variability among DfrB enzymes is a result of rapid evolution in response to drug pressure, the relative abundance of synonymous and non-synonymous substitutions were compared. The average synonymous substitution rate \((d_s)\) for the *dfr*B genes is \(0.444\pm0.036\), while the nonsynonymous rate \((d_a)\) is \(0.112\pm0.018\). These values suggest that the different mutations observed for the *dfr*B genes have not been guided by positive selection, as a result of trimethoprim pressure, but correspond to neutral or purifying selection over a long period of time.

The results of the phylogenetic analysis of the DfrB family I performed show no clear evolutionary relationship among the *dfr*B genes (Figure 3-5). It is not possible to establish a clear temporal connection among genes, as the level of sequence variability does not correlate with the temporal isolation of the different DfrB family members. Some members that were isolated more than thirty years apart (e.g. *dfr*B1 and DfrB6), share more homology than those isolated the same year (e.g. *dfr*B1 and *dfr*B2). This, combined with the suggestion that there has been no apparent positive selection within the *dfr*B family, makes the analysis of a possible evolutionary connection even more difficult.

![Phylogenetic tree of the dfrB gene family.](image)

The minimum evolution algorithm, as implemented in MEGA 3.1 [521], was used to construct the tree. Values in green indicate the robustness of each node (branch) as predicted by a bootstrapping of 500 replicates. The reported year of initial isolation for each of the genes is

---

74
Chapter 3  DfrB dihydrofolate reductase

given in brackets. The results show there is no clear correlation between the position of the dfrB member in the tree and the year of isolation.

My sequence alignment studies of the complete dfrB cassettes indicate that whereas the non-coding region that precedes the ORF is the least conserved, the attC recombination site is highly conserved (Figure 3-6), with identity levels above 83%, higher than those of the ORFs.

Figure 3-6 attC recombination site.

Comparison of hundred of cassettes [557,558] showed that the most conserved regions of several attC recombination sites are located at both ends, a 7-bp region with the consensus sequence RYYYAAC at the 5' end and a GTTRRRY sequence at the 3' end (where R represents a purine and Y a pyrimidine). These are inverted repeated sequences, with a variable region between them, which is often also an imperfect inverted repeat, giving them the potential to form a stem loop structure [559] (Figure 3-6B). This structure, as clearly suggested by the degree of sequence conservation, is expected to play a main role in cassette stability and transferability [560].
Although the \textit{attC} recombination sites of different cassettes found within integrons in mobile elements show variable sequences, the \textit{attC} sites of cassettes within chromosomally located integrons are highly conserved and seem to be species-specific [534]. The fact that all \textit{dfrB} cassettes share the same \textit{attC} site suggests that there must have been a common ancestor and that the ORF was probably originally recruited by a single bacterial species. As no other cassettes with the same \textit{attC} sequence have been reported so far, nothing can be hypothesised regarding the probable host in which this protein was originally sequestered.

The 5’ non-coding regions of the gene cassettes located upstream of the ORF present very little homology (48 to 80%) compared with that of the rest of the cassette. The mutation frequency along the complete cassette shows a distinctive pattern (Figure 3-7), with most mutations accumulating towards the 5’ non-coding region of the cassette, and the coding region corresponding to the N-terminal region of the protein.

![Figure 3-7 Nucleotide variability along the \textit{dfrB} cassette.](image)

After alignment of the \textit{dfrB} cassettes, the nucleotide variability at each position was determined. This is shown above a schematic representation of the cassette as the number of different bases, with 0 indicating conservation across all cassettes. It can be clearly seen that most of the variation is located at the 5’ non-coding region of the cassette, the coding region corresponding to the N-terminal domain of the protein, and the non-coding region between the ORF and the \textit{attC} recombination site. The most conserved part is the \textit{attC} recombination site, followed by the ORF region coding for the C-terminal domain of the DfrB protein.
The most conserved region of the *dfrB* cassettes corresponds to the *attC* recombination site at the 3’ end of the cassette. This heterogeneity in the mutational pattern of the gene cassette highlights those regions important for the propagation of the cassette -The *attC* recombination site and the ORF coding for an active enzyme.

### 3.5 Structural analysis of the DfrB proteins

R67 DHFR, a DfrB1 enzyme, is a homotetrameric protein with a highly symmetrical D2 (point group 222) structure, where four monomers contribute to the assembly of a single central active-site pore that traverses the tetramer (Figure 3-8). The openings of this pore are ellipsoidal, with a major axis of 24 Å and a minor radius of 18 Å, and it contracts towards the centre reaching 12 Å and 9 Å in the middle of the active site. This hourglass-like structure presents four equivalent binding sites, which, due to steric constraints, can accommodate up to two ligands at the same time, with R67•DHF•NADPH being the productive ternary complex [502].

#### 3.5.1 Methods

The crystal structure of the *apo* R67 DHFR protein (PDB code 1VIE) was used to search for homologues proteins. The DALI server [561], was used to explore the PDB database for structural homologues. Searches were done with the monomeric, dimeric and tetrameric structure. Several graphical programs, including Rasmol [562], VMD [563] and Pymol [564] were used to analyse and compare different structures.

#### 3.5.2 SH3 Domain

Each subunit of the DfrB enzyme presents a typical SH3 (Src Homology 3) barrel structure, where five β-strands form two orthogonal anti-parallel β-sheets of three strands each, with the third strand shared by the two β-sheets (Figure 3-9A). This is a compact barrel structure with a hydrophobic core in which the strands are connected by three loops, which in turn contribute to the oligomerization of the subunits.
Figure 3-8 Structure of R67 DHFR (1VIE).

The four monomers that form the active enzyme are depicted as differently coloured cartoons (chain A, blue; chain B, red; chain C, orange; and chain D, green), within a surface representation of the tetramer. (A) Top view of the structure: the active-site pore is located at the centre of the complex, where four monomers contribute to the binding of the ligands. (B) Side view of the enzyme: the surface representation of the molecule has been limited to half of the tetrameric structure to facilitate visualization of the double funnel-like central active-site pore. This presents elliptical openings with a major radius of 24 Å and a minor radius of 18 Å and contracts towards the centre reaching 12 Å and 9 Å, respectively, in the middle of the active site. (C) Schematic representation of the tetrameric structure of R67 DHFR and the reactants dihydrofolate (DHF) and NADPH. Amino acids suggested by experiment to be important for the binding of the ligands have been roughly positioned along the active-site pore. Each side of the pore, A-D or C-B, presents two sets of equivalent residues and is expected to accommodate one of the two ligands DHF or NADPH.

SH3 is a small domain (typically 60 residues) present in a wide variety of proteins, mainly eukaryotic, that participate in cell-cell communication and signal transduction pathways [565,566]. These proteins take part in many different events including enzymatic regulation, modulation of concentration and distribution of components of signalling pathways and assembly of protein
complexes. It functions primarily as an auxiliary domain that mediates protein-protein interactions by binding short proline-rich sequences. But is has never been found before as the unique domain of a protein. Although other enzymes have been reported to contain the SH3 fold [567-569], DfrB is the first case in which it constitutes the catalytic active site itself, rather than having a purely structural or binding role.

Figure 3-9 Monomers association in R67 DHFR.
Cartoon representation of the crystal structure of R67 DHFR (PDB code 1VIE), a DfrB1 protein. (A) SH3 domain of the monomeric unit. (B) Dimeric complex, two monomers interact with each other forming a third β-barrel in the interface. (C) Tetrameric and active structure of the enzyme. It presents a highly symmetrical central active-site cavity, with four equivalent binding sites where both substrate dihydrofolate and cofactor NADPH bind cooperatively.

Although there is substantial sequence diversity among SH3 domains, structural conservation is remarkable, with a superposition of the β-sheet core of different
SH3 domains giving an RMSD (root mean square deviation) of less than 2 Å [570]. The low sequence similarity among members of the SH3-fold group has led to two different hypotheses regarding their origin: an early horizontal gene transfer between eukaryotes and bacteria, or that the domain evolved in bacteria and was subsequently transferred to eukaryotes during mitochondrial endosymbiosis [571]. The fact that few prokaryotic proteins contain this domain, and that no proteins have been shown to use it for catalytic purposes, support the idea that the DfrB protein family is a novel kind of enzyme.

3.5.3 Structural Homologues

Although the structural resemblance between DfrB DHFR and other SH3 domains was noticed early, it was suggested that the lack of sequence similarity between them did not support a possible evolutionary relationship. The structural similarity was said to be the result of the limited number of possible three-dimensional configurations that small peptides can adopt [504,505]. However, after a careful analysis of the sequence and structure of DfrB DHFR in comparison with other SH3-domain proteins, I found that despite the apparent lack of sequence similarity there are a number of recognizable positions for which particular kinds of residues are needed to support proper folding and stabilization of the SH3 domain.

I found several structural homologues of the monomer domain of the DfrB DHFR protein within the PDB database. Comparisons of several SH3 domains showed that one of the main differences between DfrB DHFR and other SH3-like proteins is the length of the RT-Scr loop connecting β-strands a and b. This usually long loop plays a major role in the typical protein-protein interactions associated with the SH3 domain, as it participates in the binding of PXXP-motifs (P representing a proline and X any amino acid) of target proteins during signal transduction. While the RT-Scr loop seems to be the least variable of the loops of the SH3 domains [570], DfrB DHFR is not the only structure that presents a very short RT-Scr loop. This special group includes, among others, the tudor domain of the human survival motor neuron protein (1G5V [572]), the *E. coli* biotin holoenzyme synthetase/bio repressor (1BIA [573]), the C-terminal domain of the repressor protein KorB of *E. coli* (PDB code 1GQ [574]), and the DNA binding
domain of HIV-1 integrase (PDB code 1IHV [569]). Superposition of the aligned Cα atoms of these domains with that of DfrB DHFR gave RMS deviations of 1.9 to 2.8 Å (Figure 3-10).

![Figure 3-10 Superposition of SH3 domain structural homologues.](image)

Cα trace representation of 5 superimposed SH3 domains. The superposition of Cα atoms was done using the DALI server [561]. Front (A) and top (B) view of DfrB DHFR (1VIE, thick blue), the human survival motor neuron protein (1G5V, grey), the E. coli biotin holoenzyme synthetase/bio repressor (1BIA, magenta), the C-terminal domain of the repressor protein KorB of E. coli (1IGQ, green), and the DNA binding domain of HIV-1 integrase (1IHV, orange).

### 3.5.4 DfrB and HIV-1 integrase

The uniqueness of the DfrB enzymes extends beyond the folding of the SH3 monomers to the interesting structure of the catalytically active tetramer. During the oligomerization process two monomers form an initial dimeric complex [575] (Figure 3-9B). Whereas other SH3 domains have been shown to form homodimers [574,576,577], searches for structural homologues showed that only the C-terminal DNA-binding domain of HIV-1 integrase (IN-DBD) is structurally equivalent [569]. The two monomers come together sharing three β-strands each, and forming a compact β-barrel at the dimer interface (Figure 3-9B). A least-square superposition of the Cα atoms of 30 core residues, including all amino acids from the β-strands and the 3₁₀-helix, gave an RMSD of 1.1 Å for the monomers and 2.4 Å for the dimers (Figure 3-11). This structural similarity is reflected in the conserved positions of residues that form the
hydrophobic core as well as the positions of those residues that participate in the
dimer interface.

![Figure 3-11 R67 DHFR and HIV-1 integrase.](image)

Least-square superposition of Cα atoms of 30 core residues of the monomeric (A) and dimeric (B) SH3 domain of the DfrB protein (1VIE, blue) and the C-terminal DNA-binding domain of HIV-1 integrase (1IHV, orange). The RMSD values were 1.1 Å for the monomers and 2.4 Å for the dimers.

It has been suggested that the saddle-shaped groove of the IN-DBD dimer could be suitable for DNA binding [569,578]. In the case of DfrB, two dimers interact with each other, mainly through residues of the Distal and the short RT-Src loops, to form the final active tetrameric structure in which two opposed saddle-shaped grooves form the central active-site pore where the ligands bind (Figure 3-9C).

Although there is no obvious similarity between R67 DHFR and IN-DBD from sequence comparisons, it is clear that there is an important structural resemblance. Not only do they show a similar dimer interface, but also the nature and position of the residues that form the hydrophobic core and those that participate in the dimer interface are highly conserved. It is not clear if this is just the result of convergent evolution, where specific residues have been selected to facilitate the formation of an SH3 domain and its dimerization, or if both systems have derived from a common ancestor.
3.6 Compilation of Mutational Analyses

Extensive mutational analyses have been carried out on R67 DHFR. Point mutations \([510,545,548,579-581]\), combinatorial exploration of the catalytic site \([582]\) and \textit{in vitro} evolution studies \([583]\) have shown R67 DHFR to be a robust enzyme (Figure 3-12). Several studies carried out in Howell’s group on R67 DHFR led to the identification of some important residues that participate in the binding of the ligands DHF and NADPH and influence the overall reaction. These include K32, Q67, I68 and Y69. However, as may be seen from Figure 3-12, mutations for all these positions, which do not alter the capacity of the enzyme for conferring TMP-resistance, have been found.

Figure 3-12 Compilation of mutations observed for active DfrB enzymes.

Single and multiple mutations that result in a TMP-resistant phenotype were compiled from reported mutational analyses \([510,545,548,579-583]\). For each residue, a list of all alternative amino acids found from the collection of mutants is shown below the original DfrB1 sequence. Note that the exact sequence of each mutant is not given, and that not all possible combinations of multiple mutations have been studied. Whereas naturally occurring enzymes of the DfrB family present most of their variability within the N-terminal region (see Figure 3-4), experimentally generated mutants show a wider distribution. Even those residues suggested to be catalytically important (shaded in grey) can be substituted without loss of activity.

In a recent work, Schmitzer \textit{et al.} \([582]\) performed a combinatorial exploration of the catalytic region by mutating all 16 active-site residues, four residues per monomer: Val66, Gln67, Ile68 and Tyr69. Within their library of mutants they found three variants, two with three mutations (V66S-Q67K-Y69E and V66I-Q67N-I68R) and one with four mutations (V66G-Q67E-I68L-Y69H) which presented wild-type like enzyme kinetic properties. As no side chain previously thought necessary \([505,545-548]\) was conserved among the selected mutants, it was concluded that the catalysis of the reaction must be controlled by the
appropriate relative positioning of the substrate and the cofactor within the active site, and that this requires little, if any, participation of specific residues.

*In vitro* evolution studies [583] showed that, on average, one nucleotide change out of two or three does not disrupt the capacity of the DfrB enzyme to confer TMP resistance. Although most mutations accumulate in the N-terminal region, if the first 18 residues not required for catalytic activity are ignored, 82% of the amino acids can be replaced, up to 14% of them at the same time [583], without lost of activity *in vivo*. However, the number of mutations that have little or no effect on the catalytic properties of the enzyme is much smaller if measured *in vitro*. These results led to the suggestion that a high concentration of inactive or weakly active dimers inside the bacterial cell, combined with the presence of substrate molecules, could lead to resistance through the transient assembly of reactive ternary complexes [579]. This capacity further highlights the functional flexibility of the DfrB protein.

Although the N-terminal region appears unimportant for catalytic activity, it does seem to influence enzyme stability *in vivo*. A truncated version of the enzyme, although active *in vitro*, failed to produce TMP-resistant strains [503]. Furthermore, *in vitro* evolution analysis demonstrated that the N-terminal region was able to confer stability or folding efficiency to mutants regardless of its sequence [583]. Therefore, it has been suggested that the N-terminal region could act by protecting the enzyme from proteolytic degradation, by increasing the stability of the mRNA, or by facilitating the folding process of the monomer or the formation of the active tetramer.

The symmetric nature of the homotetrameric structure limits the mutational patterns that can be analysed, or even naturally selected. The alteration of a single residue in the primary sequence will result in four mutations per active tetramer, affecting the binding of both ligands, as well as the quaternary structure of the protein. Thus, many residues originally identified as important for ligand binding and catalysis (Lys32, Gln67, Ile68 and Tyr69 [505,545-548]) also seem crucial from a structural point of view. While Tyr69 is part of the core structure of the monomer, and its position is usually occupied by aromatic residues in SH3 domains [570], Val66 and Ile68 are positioned at the dimer
interface, and Gln67 could contribute to the stabilization of the homotetrameric complex. Therefore, the mutation of any of these residues might affect the catalytic activity as well as the active structure of the enzyme. This is another atypical characteristic of DfrB DHFR; typical enzymes do not usually have residues that are both structurally and catalytically important, as this severely constrains the evolutionary flexibility of the protein.

3.7 R67 DHFR and ligand binding

As previously mentioned, the catalytically active enzyme has a highly symmetrical structure with a single active-site pore that traverses the tetramer. The active site presents an hourglass shape with wide openings and a narrow central region. There are no separate binding sites for substrate (DHF) and cofactor (NADPH), but four equivalent binding sites per tetramer. The size and shape of the cavity constrain the number of ligands that can be positioned simultaneously: two substrates (DfrB•DHF•DHF), two cofactors (DfrB•NADPH•NADPH), or one of each (DfrB•DHF•NADPH) can be accommodated within the active-site pore [502].

It is known that the same residues, located in symmetrically equivalent positions, interact with both ligands [507,548], and that inter-ligand cooperativity plays an important role in the binding process, with NADPH facilitating the binding of DHF [502]. However, despite much experimental work, including X-ray crystallography [505], interligand nuclear Overhauser effects (NOEs) [513] and other NMR studies [547] (Table 3-3 and Figure 3-13) a clear picture of the reactive ternary complex DfrB•DHF•NADPH has not emerged yet. This lack of success might be due, in part, to the mobility of the ligands within the spacious pore. Despite interligand NOE experiments [513] suggesting that while the pterin and nicotinamide rings of the ligands are located close to each other near the centre of the pore, and the rest of the molecules extend towards opposite ends, there is still not enough information to accurately position the ligands within the active site.

The symmetric nature of the active site (Figure 3-8) results in two equivalent residues from different monomers on each side of the pore, each of them capable of establishing similar interactions with the ligands. The possibility of non-
unique binding modes is further increased by the flexibility of the ligands, which can adopt multiple conformational spaces within the spacious central pore in agreement with empirical observations. Therefore, in order to obtain a reliable structure of the reactive complex, I employed computational methods to thoroughly explore the conformational space of the ligands within the enzymatic cavity, and select the most stable conformations compatible with the experimental results (Table 3-3 and Figure 3-13).

![Figure 3-13 Schematic representation of the reactants DHF and NADPH and the interactions they establish with the enzyme R67 DHFR as observed experimentally.](image)

The first attempt to generate a theoretical model of the complex R67•DHF•NADPH was carried out by Howell et al. [507]. They used the docking programs DOCK [165] and SLIDE [206] to produce structures of a R67•folate•NMN (nicotinamide mononucleotide moiety of NADPH) complex together with a R67•pterin•NADPH complex [507]. A structure for the complete reactive complex R67•DHF•NADPH was not reported. Moreover, the partial models were not analysed for stability nor was their reliability tested with other methodologies (e.g. MD simulations).
Table 3-3 Experimental observations pertinent to the conformations adopted by the ligands DHF/folate and NADPH/NADP+ within the active site of R67 DHFR.

<table>
<thead>
<tr>
<th>Ligand / residue</th>
<th>Observation</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Folate/DHF</strong></td>
<td>Only the position of the pterin ring could be determined crystallographically. The O4 atom appears H-bonded to the backbone of I68.</td>
<td>X-ray crystallography [505].</td>
</tr>
<tr>
<td></td>
<td>The glutamate moiety does not have a defined binding orientation.</td>
<td>Interligand Overhauser Effects [513].</td>
</tr>
<tr>
<td></td>
<td>It is involved in at least one ionic interaction with the enzyme, probably K32, at low salt concentrations.</td>
<td>Isothermal titration calorimetry, ionic strength effects [545].</td>
</tr>
<tr>
<td></td>
<td>Important ILOE peaks connecting H9 (folate) with H4 and H5 of NADP+.</td>
<td>Interligand Overhauser Effects [513].</td>
</tr>
<tr>
<td></td>
<td>During the reduction the hydrogen atom is transferred to the C6 si face of the ring</td>
<td><em>In vivo</em> growth assay of <em>E. coli</em> strain C600 [512].</td>
</tr>
<tr>
<td><strong>NADPH/NADP+</strong></td>
<td>Adopts an extended conformation when bound to the enzyme.</td>
<td>Interligand Overhauser effects [513].</td>
</tr>
<tr>
<td></td>
<td>The ribonicotinamide bond adopts a <em>syn</em> conformation while the adenosine moiety adopts an <em>anti</em> conformation.</td>
<td>NMR, Interligand Overhauser Effects [513,584].</td>
</tr>
<tr>
<td></td>
<td>The diphosphate and the phosphate groups participate in two ionic interactions with K32 residues.</td>
<td>Isothermal titration calorimetry, ionic strength effects, fluorescence quenching [545].</td>
</tr>
<tr>
<td></td>
<td>The diphosphate subunit may interact with Y69.</td>
<td>NMR [547].</td>
</tr>
<tr>
<td></td>
<td>Its binding causes large chemical shifts in the signal of the side chain of Q67.</td>
<td>NMR [547].</td>
</tr>
<tr>
<td></td>
<td>Its binding to the enzyme is accompanied by minor changes in the backbone.</td>
<td>NMR [547].</td>
</tr>
<tr>
<td></td>
<td>The <em>pro-R</em> H4 hydrogen of the nicotinamide ring is the one transferred during the reaction.</td>
<td>NMR [512].</td>
</tr>
<tr>
<td><strong>K32</strong></td>
<td>Interacts with both ligands. May confer a general positive potential to the pore.</td>
<td>Mutational analysis, isothermal titration calorimetry, ionic strength effects [545,546]. Electrostatic potential calculations [507].</td>
</tr>
<tr>
<td><strong>Q67</strong></td>
<td>Appears interacting in pairs with a symmetry related Q67 residue. Undergoes large chemical shifts upon NADPH binding.</td>
<td>X-ray crystallography [505], NMR [547].</td>
</tr>
<tr>
<td><strong>I68</strong></td>
<td>Its backbone forms H-bond interactions with the carbonyl group of the pterin ring.</td>
<td>X-ray crystallography [505].</td>
</tr>
<tr>
<td><strong>Y69</strong></td>
<td>Its hydroxyl group may interact with both ligands.</td>
<td>Mutational analysis [546,548].</td>
</tr>
</tbody>
</table>
Many new experimental studies on R67 DHFR [545-547, 581, 585, 586] have been published since the construction of this partial model, providing further insights into the reactive complex and the mechanism of R67 DHFR. Some of this new data, such as the interaction between the hydroxyl group of Y69 and NADPH observed in mutational analysis [25], is not in agreement with the previous docking results [507] and, therefore, the model required re-examination.

The generation of an updated, more rigorously tested, model of the reactive complex R67•DHF•NADPH was the first major goal of my thesis project. I used the docking programs AutoDock [166] and FlexX [587] to explore the conformational space of the ligands, protonated DHF (DHFH+) and NADPH, inside the active site of R67 DHFR. Different final poses were further ranked using a consensus scoring analysis and three different scoring functions. Finally, the best conformations were submitted to MD simulations, to test their stability over time. The results of the MD simulations provided further insight regarding the behaviour of the ligands within the protein environment, the flexibility of key residues involved in protein-ligand interactions, and the overall strategy used by the enzyme to accommodate the reactants within its spacious active-site pore.

### 3.7.1 Methods

#### 3.7.1.1 Docking

R67 DHFR exhibits a unique active-site pore where both reactant and cofactor bind to equivalent sites. This feature, together with the size and symmetry of the cavity and the importance of cooperativity between ligands for the formation of a productive ternary complex [502] made it necessary to adopt non-conventional docking approaches.

Although there have been several recent reports comparing various docking programs and scoring functions [141, 159-161, 218, 235, 588], the results of these suggest that the performance of different docking tools is very dependent on the target protein, and that comparative analysis may lead to different results depending on the properties examined (quality of the top-ranked poses, quality
of all poses, predicted binding energies, docking time, etc). As most of these studies focus their attention on the screening of virtual libraries, the settings of the programs are optimised for fast docking rather than docking quality. Furthermore, the systems analysed usually include small-sized molecules and proteins with simple active sites, excluding difficult cases such as R67 DHFR. The docking programs AutoDock [166], which implements a genetic algorithm, and FlexX [587], with an incremental construction algorithm, were selected to carry out the docking analysis. Both programs have been used successfully to dock several ligands [587, 589-592], including methotrexate, a folate analogue, to the active site of a type 1 DHFR [166, 167, 587]. AutoDock has been shown to be adequate for predicting multiple binding modes [593], while FlexX is claimed to be particularly good for studies where the active site is predominantly lipophilic but features polar groups that must form specific interactions with the ligands [161], as is the case for R67 DHFR.

Three different docking strategies were pursued, including: a sequential docking, where one molecule/moiety was docked after another; a superligand approach, where both DHFH+ and NADPH were covalently bonded and docked as a single entity; and a manual placement, were the ligands were manually guided within the active-site pore (Figure 3-14).

3.7.1.1.1 Structure preparation

There are only two R67 DHFR crystal structures available, one of the apo enzyme (1VIE), and one of a R67•folate complex (1VIF) [505]. A comparison of these structures shows minimal conformational change upon ligand binding (backbone RMSD 0.02 Å, side chain RMSD 0.05 Å). Therefore, the crystal structure of R67 DHFR complexed with folate (1VIF) was considered to be an appropriate starting point for the building of the complexes without further modifications. As the 1VIF structure provided two alternative positions of the pterin ring, the one with the si face of the C6 centre exposed was selected (R67•pte), based on the stereochemistry of the reaction [512]. Structural coordinates for the ligands, protonated dihydrofolate (DHFH+) and NADPH, were obtained by modification of the crystal-structure coordinates of folate and NADP+ from 1RA2 [100]. A “superligand” complex was constructed by
covalently connecting the reacting H atom to both C4 of NADPH and C6 of DHFH+.

Figure 3-14 Docking sequence performed with AutoDock [166] (A) and FlexX [587] (B). Red crossed arrows indicate unsuccessful attempts, dotted arrows denote doubtful results, and green arrows show successful docking. Those complexes shown in blue correspond to the structures that were further analysed, including complexes adt_109 and adt_158 obtained with AutoDock and complexes flx_d1n2, flx_n1d2 and flx_super generated with FlexX. R67-pte, R67 DHFR including the crystal position of the pterin ring of folate; NMN, nicotinamide-ribose-phosphate moiety of NADPH; pte, pterin ring. a This corresponds to a manually guided docking of DHFH+ into R67 DHFR, taking the crystal position of the pterin ring as reference.

3.7.1.1.2 AutoDock

AutoDock 3.0.5 [166], which uses a stochastic-search based algorithm, was employed for the generation of ternary complexes. Hydrogens and charges were
added to the ligands using InsightII [594] (CFF91 force field). Autotors, part of the AutoDock docking package, was used to define the active torsions. Kollman charges were assigned to the enzyme using the AutoDock tools program, and the Addsol module was used to generate the volume parameters. Affinity grids of 60x90x60 points separated by 0.375 Å were produced around the active-site pore using AutoGrid. The Lennard-Jones and H-bonding potentials supplied with the program were used to calculate the interaction grids.

The Lamarckian genetic algorithm (LGA) was used for the conformational search. Each LGA job consisted of 200 runs and the number of iterations of the pseudo-Solis and Wets local search was set to 300. The initial population was between 200 and 250 structures, while the maximum number of energy evaluations and generations was 10 million; hence, the energy evaluations were the limiting factor. These values were chosen based on the systematic study done by Hetényi and van der Spoel on the influence of different parameters in the docking of peptides to proteins, without prior knowledge of the binding site [595]. The rest of the parameters were set to default values. The final structures were clustered and ranked according to the native AutoDock scoring function.

### 3.7.1.1.3 FlexX

FlexX 1.13.5 [587], which implements an incremental construction algorithm, was the second docking program used. The active-site region was defined by the residues K32, A36, Y46, G64, S65, V66, Q67, I68 and Y69. Both automatic and manual anchor-base selections were performed. The placement of the anchor bases was done using the triangle hashing technique and the manual base placement; in the last case the crystal position of the pterin ring, R67•pte, was used as the reference structure. The final poses were clustered and scored using the basic scoring function of FlexX (similar to the Böhm function [232]). Use of the particle positioning option, to model possible water molecules important for the binding process [175], was also employed, but no important differences were found (results not shown).
3.7.1.4 Manual Docking

InsightII [594] was used to manipulate one of the ternary structures obtained with AutoDock in order to change the conformations of DHFH+ and NADPH ligands to those in conformity with all the experimentally observed properties (see below). The final structures were optimised by a short energy minimization cycle (100 steps of steepest-descent) with GROMACS [261,596].

3.7.1.5 Analysis of the structures

The generation of a series of ternary structures through the use of different methods produced a diverse set of complexes sufficient for further studies. In all cases, the final poses were analysed and selected taking into account all the available empirical information. Only the best-ranked structures of each docking approach that fulfilled most of the experimentally observed properties, with particular emphasis on those closely related to the hydride transfer step (e.g. distance between the reacting centres), were further studied (Figure 3-14). The constraints used for the analysis and selection of the best poses included:

- C6(pte)-C6(nic) distance: only structures with the reacting centres within 4 Å were further analysed.

- C6(pte)-H-C4(nic) angle: guided by the exo model for the reactant state proposed by Castillo et al. [11] The angle was required to be larger than 100°.

- C8A(pte)-C6(pte)-C4(nic)-N1(nic) dihedral: the endo conformation of the rings required dihedral angles between -90° and 90°.

- anti conformation of the adenosine glycosidic angle and syn conformation of the ribonicotinamide glycosidic angle of NADPH according to transfer NOE and interligand NOE (ILOE) studies [513,584].

- H-bond interactions between the phosphate groups of NADPH and K32 [545,581].
Using the coordinates for some of the published docked complexes for folate and NADPH [507], kindly provided by Prof E. Howell (University of Tennessee), a ternary complex structure was generated. The coordinates of folate were taken from the highest scoring R67•folate•NMN complex generated by DOCK, and those of NADPH, from an R67•pte•NADPH complex consistent with NMR data. These were combined to generate the final R67•DHFH+•NADPH complex (howH), after prior modifications of the protonation and reduction state of the pterin ring of folate.

### 3.7.1.1.6 Multiple Scoring

The selected ternary complexes were ranked using three different scoring functions: AutoDock native scoring function [166], Böhm-like scoring function [232] and DrugScore function [233]. While AutoDock and the Böhm-like scoring functions belong to the so-called empirical group, where the binding free energy is calculated using a summation of terms fit to experimental binding constants, DrugScore is a knowledge-based potential, where distance-dependent interactions obtained from crystallographic data are used to evaluate the poses.

Due to the presence of two different ligands in the reactive complex, two basic approaches were used for the scoring of the structures. A serial score, where the interaction between one of the ligands (DHFH+ or NADPH) and the protein (R67 DHFR) was initially evaluated, followed by the scoring of the second ligand binding into the previously evaluated binary complex (R67•DHFH+↔NADPH or R67•NADPH↔DHFH+), the sum of these two values being the overall score. And an independent score, where the interaction energy of each ligand with the protein was assessed independently (R67↔DHFH+, R67↔NADPH), and the sum of these values was considered as the total energy, i.e. interactions between the ligands were neglected. Therefore, the contribution to the total energy given by the interaction between the ligands was considered to be the difference between the serial and the independent scores.
3.7.1.7 Consensus scoring

The evaluation of the interaction energy between a docked molecule and its receptor is not precise, as the calculation often involves several approximations and the use of empirically or theoretically parameterised functions. Therefore, there is not a “best” scoring function, because different algorithms will behave differently in different situations. The results often depend on the nature of the complex itself, the type of scoring function, the parameters it uses to evaluate different poses, the training data set used during the parameterisation procedure, etc. Therefore, instead of relying on only one function, several algorithms can be used and their results combined into a single consensus score [234,597].

The results of the three different scoring functions were combined into four different consensus scores. The simplest combination was the addition of the three ranks, giving the sum rank. The worst-best rank involved assigning each pose the second-worst rank value, while the sum of all but the worst score provided the so-called deprecated rank-sum. Finally, we performed a consensus score (cons-score) where the different poses were given one vote in favour if their score was within the best half (top four) or best third (top three) structures out of the eight complexes according to each function. The final cons-score was the combination of points for the best-third and the best-half criteria.

3.7.1.2 MD simulations

After performing the consensus score, all R67•DHFH++NADPH complexes were further analysed using MD simulations. The GROMACS simulation package [261,596] was used to study the evolution of the complexes with time. The simulations were carried out using the OPLS-AA force field [598,599]. Charges and parameters for the ligands were assigned by fragment comparison with homologous parameterised structures, except for the reacting rings which were calculated at the B3LYP/6-31G* level using N5-protonated 6-methyl-dihydropterin and 1-methyl-nicotinamide fragments as models.

The different structures were solvated in a triclinic box of SPC water molecules [600], leaving a minimum distance of 8.5 Å between the complex and the edge of the box. The final charge of the systems was –4; no counterions were added.
as they were seen to form strong ionic interactions with the charged ligands during MD simulations (results not shown), perturbing their interactions with the protein. The Particle-Mesh Ewald (PME) method was used to treat long-range electrostatic interactions [601]. The solvated structures were submitted to a short energy minimization process (100 steps of steepest-descent) and further optimised by performing a 100 ps simulation of water molecules only. Finally, the complete systems were simulated for 4 ns each. All atoms were treated explicitly and the SHAKE algorithm was used to constrain the length of covalent bonds between hydrogen and heavy atoms. The time step was 0.001 ps and the Coulomb and van der Waals cutoffs were set to 1.2 nm. The simulations were performed at constant volume and temperature. Initial velocities were assigned based on a Maxwell distribution at 300 K. The protein and ligands, and the water molecules were coupled separately to a bath at 300 K using the Nosé-Hoover algorithm and a coupling time of 0.1 ps. The centre of mass motion was removed in every step, while the neighbour list update was done every 10 steps.

An initial 1 ns simulation of the apo enzyme was carried out in order to test the stability of the system under the chosen conditions. The different trajectories were analysed using several auxiliary programs provided within the GROMACS package. These included \texttt{g\_bond} for the calculation of distances, \texttt{g\_angle} for the measurement of angles and dihedrals, and \texttt{g\_hbond} for the H-bond interactions between ligands and protein. In the last case, all the possible H donors and H acceptors in both the ligands and the protein were taken into account, and a distance of 3 Å and an angle of 60° were used as the cutoff for the interaction.

\subsection*{3.7.2 Results}

\subsubsection*{3.7.2.1 Docking}

The main challenge of this docking study was to position two different molecules (DHFH+ and NADPH) within one common central pore, taking into account that the interactions between the molecules themselves may be critical for the overall binding process [502]. Therefore, two different strategies were developed to dock the ligands DHFH+ and NADPH into R67: a step-by-step docking procedure, where one molecule/moiety was docked after another and a
superligand approach, where the two molecules were covalently bound and docked as a single unit.

The ternary complexes R67•DHFH+•NADPH were constructed using the docking programs AutoDock and FlexX, as well as a manual approach. Assessment of the final poses was judged against the empirically observed properties (Table 3-3). Only those structures that presented reasonable interactions between the protein and the ligands, and between the ligands themselves, were further analysed.

### 3.7.2.1.1 AutoDock docking

Only a few of the docking approaches tried with AutoDock worked (Figure 3-14). No sensible structures were obtained when docking DHFH+ or NADPH into the apo enzyme. It was necessary to initially take into account the crystallographic position of the pterin ring in the enzyme structure (R67•pte) in order to obtain some reasonable structures. As the docking of the complete structure of NADPH into R67•pte did not produce any sensible conformations we proceeded to dock a truncated version of NADPH, a nicotinamide-ribose-Pi moiety (NMN). This approach gave good results: 25% of 200 final structures clustered in the largest group (RMSD smaller than 1 Å) with the lowest docking energy, all of them presenting reasonable conformations with a C6(pte)-C4(nic) distance shorter than 4 Å (Figure 3-15A). After selecting the best-ranked R67-pte•NMN complex, the pterin ring was removed and DHFH+ was docked. Although half of the 200 final structures presented the pterin ring of DHFH+ in close proximity to the crystallographic position, the positioning of the para-amino benzoyl glutamic acid tail (pABA-Glu) showed significant variability, in agreement with published experimental results [513,602] and previous docking studies [507] (Figure 3-15B).
Figure 3-15 Relative positioning of ligands in the final docked structures.

(A) Cluster of the highest scoring nicotinamide-ribose-Pi moiety conformations docked into R67•pte using AutoDock. This corresponds to the largest cluster of structures (RMSD smaller than 1 Å), which also exhibit the lowest average energy. The top scoring pose and the pterin ring are drawn in tubular form, whereas the remainder of the nicotinamide-ribose-Pi subunits are represented as wireframe. Only polar hydrogen atoms considered during the docking process are shown. (B) Orientations of different N5-protonated dihydrofolate conformers obtained after the docking of the substrate into R67•NMN using AutoDock. All structures present the pterin ring in a position close to the crystallographic one. The para-amino-benzoyl-glutamic tail adopts two main conformations depending on the K32 residue with which it interacts. A representative of each of these two main groups is shown in tubular form, with the remainder in wireframe. (C) Conformation of 8 selected complexes of N5-protonated dihydrofolate and NADPH within the active site of R67 DHFR. Complexes adt_109 (red), flx_n1d2 (blue) and howH (orange) appear in tubular form, with the others in wireframe (adt_158: green, flx_d1n2: yellow, flx_super: violet, man_react: magenta, and man_TS: cyan). The active site of the protein has been included as a trimmed surface, where the position of key amino acids has been highlighted in different colours, K32: red, Q67: green, I68: white (centre) and Y69: blue. It can be seen that while the pterin ring of DHFH+ and the nicotinamide ring of NADPH are located in similar positions (except flx_super), the tails extend towards opposite ends adopting several conformations.
The pABA-Glu tail adopted two major conformations, determined by the interaction of the glutamic moiety of DHFH+ with one of the two K32 residues present in the half-pore region occupied by the reactant. There was an extended conformation, where both the pterin ring and the Glu moiety were interacting with the same dimer-dimer interface (D/C or B/A); and a bent one, where the pterin ring of DHFH+ was interacting with the interface between chains D and C (or B and A) while the glutamic moiety was H bonded to the K32 residue of the other interface, B/A (or D/C).

Raman studies of DHF in R67 [602] showed a split signal of the C6=N5 stretching mode. It was suggested that this was the result of two possible binding modes, in which the substituents of the C6-R bond adopted different orientations. Also, recent asymmetric mutational analysis showed that double mutants where K32 from monomers A and B or A and C were changed to methionine exhibited similar kinetic parameters, leading to the conclusion that both topologies permitted similar interactions between K32 and DHF [546]. All these observations are in clear agreement with the docking results.

Finally, after selecting the best R67•DHFH+•NMN complex, the NMN unit was removed and NADPH was docked. This new placement provided very few structures with a reasonable overall conformation; there were no two conformers with the same positioning of the tail. These results, combined with the failed docking of the superligand molecule, suggest that the size of the docked molecules could be a limiting factor in the generation of reliable complexes using AutoDock. The ligands, especially NADPH and the superligand, have many torsional angles (18 and 27, respectively) and, therefore, the conformational space that the program has to explore might be too large to be properly handled by the LGA algorithm. On the other hand, when small molecules were used, NMN (6 active torsions) or DHFH+ (11 active torsions), there was less variability within the results, which were in good agreement with experimental data. Two of the final R67•DHFH+•NADPH complexes were chosen and selected for further analysis (see below).
3.7.2.1.2 FlexX docking

FlexX performed better than AutoDock when docking NADPH into the apo enzyme. The use of an incremental growth algorithm may have provided a better treatment of the large number of active torsions of NADPH compared with the LGA approach. However, no poses were obtained for the automatic docking of DHFH+. When the crystallographic position of the pterin ring was used for the docking of NADPH, or for the manual guidance of DHFH+, some reasonable complexes were generated. Selected structures from the guided positioning of DHFH+, 15 in total, and the automatic docking of NADPH, 6, were further used to dock the second ligand. All the final complexes obtained were systematically screened to find those that best fulfilled the experimentally determined characteristics. There was a lot of variability among the tertiary complexes and generally the most sensible poses were poorly ranked and were not part of big clusters. The docking of the superligand did not provide very clear results; very few structures conformed to experimental restrictions. The best poses were selected for further study (Figure 3-14).

Howell et al. also observed this variability among final poses in their original construction of partial complexes using the docking programs DOCK and SLIDE [507]. Therefore, we considered that the inability of the docking programs to find a unique structure is likely to be a result of the features of the system rather than deficiencies in the docking algorithms themselves.

Due to the heterogeneity of the final structures, a manual docking procedure was carried out, taking into account all the experimentally observed properties. Two different complexes (man_react and man_TS) were generated, the interactions between the rings were adjusted based on the structures published by Castillo et al. [603] for an in vacuo model of the reaction. The main difference between the two structures, apart from that of the reacting rings, was the positioning of the phosphate groups and adenine ring of NADPH.

A superposition of the best structures of each successful docking approach showed important differences in the conformations of the ligands, especially of the 2’,5’-ADP tail of NADPH (Figure 3-15C), even though all of them conformed to empirical observations. The variability among the final poses, and
the lack of more detailed experimental data, made discrimination of the best structure far from trivial. I carried out some comparisons in order to test the stability and feasibility of the eight possible ternary complexes; these comprised the two structures generated with AutoDock (adt_109 and adt_158), three with FlexX (flx_d1n2, flx_n1d2 and flx_super), two manually constructed complexes (man_react and man_TS), and the structure built from published results (howH).

3.7.2.1.3 Comparative Scoring

Three scoring algorithms were used to rank the structures: the AutoDock native scoring function [166], the Böhm-like scoring implemented in FlexX [587] and DrugScore [233,587]. As two different molecules were docked inside the same active site, two different strategies were employed to evaluate the poses: a serial scoring and an independent scoring (see section 3.7.1.1.6 for more details). In the serial scoring of the eight different complexes, the order in which the ligands were scored did not influence the final result (data not shown). Although the energy of the independent scores was higher (less negative) than that of the serial ones, the relative order of the complexes was not altered (Figure 3-16A).

The contribution to the final energy given by the interaction between the two ligands was very uniform among the different structures according to DrugScore, with average values of -14 kJ/mol, while AutoDock and the Böhm-like function predicted a greater range of interligand interactions energies. In the case of AutoDock, the values ranged from -6.6 kJ/mol for howH to -22.1 kJ/mol for adt_109, and in the case of the Böhm-like function from -4.3 kJ/mol for adt_109 to -19.5 kJ/mol for flx_n1d2. These values represent around 13% of the total docked energy in the case of DrugScore, about 20% for AutoDock, and up to a 50% in the case of the Böhm-like function. This significant variability among the different functions could be reflecting the fact that they have been optimised to estimate the interaction energies between ligands and protein and not between ligands themselves.

Overall, the three scoring functions gave different rankings (Figure 3-16A). The best correlation was observed between AutoDock and DrugScore (r = 0.82), whereas the results of the Böhm-like function were considerably different.
Figure 3-16 Scoring analysis of docked structures.

(A) Final docked energy after serial scoring of different selected poses using three different algorithms: AutoDock native scoring function [166] and DrugScore [233,587], and the Böhm-like [587] algorithm implemented in FlexX. The interaction energy of N5-protonated dihydrofolate with R67 DHFR was first computed, followed by the scoring of the NADPH moiety into the binary complex R67•DHFH+. Shaded regions represent the contribution to the final energy given by the interaction between the two ligands, calculated as the difference between the total energy of the serial and the independent scores. (B) Consensus scoring, the results of the serial scoring procedure were combined into a cons-score, rank-sum, worst-best and deprecated rank-sum. *Cons-score is the only consensus score where a bigger score corresponds to a better structure.
While AutoDock and DrugScore favoured the complexes generated with AutoDock (adt_109 and adt_158), the Böhm-like function identified the FlexX structures as the best ones (flx_n1d2 and flx_d1n2). It is interesting to note that the structure constructed from coordinates of the published partial complexes is one of the lowest scoring poses, and it is the only one to yield a positive energy with the Böhm-like scoring function.

### 3.7.2.1.4 Consensus Scoring

In order to resolve the disparities among the scoring functions, and identify the best ternary complex out of the eight selected structures, the individual scores were combined into four different consensus rankings: cons-score, rank-sum, worst-best rank and deprecated rank-sum [597] (Figure 3-16B). The results of all these combinations agreed on the overall final ranking of the structures, giving the following order of structures from best to worst: adt_109, adt_158, flx_n1d2, flx_d1n2, man_react, flx_super, howH and man_TS.

### 3.7.2.2 MD simulations

After the consensus scoring, all the complexes were further analysed by conventional MD simulations. The structures were solvated and simulated under periodic boundary conditions for 4 ns using the GROMACS MD package [261]. In order to evaluate the stability of the complexes, and the likelihood of their leading to a productive transition state, several parameters were monitored during the simulation. These included the interactions between the nicotinamide and pterin rings, such as the distance between the two reactive centres C6(pte) and C4(nic), the angle C6(pte)-H-C4(nic) and the dihedral C8A(pte)-C6(pte)-C4(nic)-N1(nic), and the H-bond interactions between the ligands and the protein.

#### 3.7.2.2.1 Movement of the protein and the ligands during the simulations

The apo enzyme was initially simulated for a period of 1 ns to test the stability of the system. The structure did not exhibit major deviations from the crystal structure reaching a global backbone RMSD of 3 Å while the active-site backbone presented a stable RMSD of 1.5 Å; similar values were obtained for
all other complexes (see Figure 3-17 as an example). The conformations of key residues, including K32, V66, Q67, I68 and Y69 (see below) were also analysed along the different trajectories. Most of these presented conformations comparable with those found in the crystal structure; the only residues that showed variable torsional angles were K32 and Q67, which participate in different H-bond interactions with the ligands, as, discussed below. The movement of Q67 upon ligand binding has been observed experimentally (Table 3-3).

Figure 3-17 Ligand mobility within the active site pore.
RMS deviations of different regions of the ligands DHFH\(^+\) (A) and NADPH (B) after superposition of the backbone of residues lining the active-site pore (K32, A36, Y46, G64, S65, V66, Q67, I68 and Y69). These values correspond to the 4 ns MD simulation of the complex howH. All simulations except flx_n1d2 and flx_super presented the same pattern.

To analyse the movements of the ligands, DHFH\(^+\) and NADPH were each divided into three different regions, for which the RMSDs were tracked during the simulations, after superposition of the active-site backbone (Figure 3-17A, B). It may be seen that while the reacting rings, PTER and NICO, exhibit deviations comparable with those of the backbone, the tails of the molecules display much greater fluctuations. Movement of the ligand regions becomes more evident towards the extremities of the molecules, GLU and RIBADEPHO,
which are situated close to the openings of the active-site cavity. The same pattern was observed for all but two simulations, flx_super and flx_n1d2, where the RMSD of the rings was comparable with that of the other regions of the ligands, indicating the instability of these conformations.

3.7.2.2 Interactions between the reacting pterin and nicotinamide rings

There is enough experimental evidence to support the idea that the disposition of the reacting rings within R67 DHFR is quite different from that found in the chromosomal enzyme. Although the pterin and the nicotinamide rings are known to adopt an \textit{endo} conformation in R67 DHFR there are no studies on the mechanism of the reaction and, therefore, there is not a clear picture of the global arrangement of the reactants. There is only one study where an \textit{endo} transition state for a model of the reaction \textit{in vacuo} was presented as an alternative to the \textit{exo} conformation found in the ubiquitous type 1 DHFR \cite{513,603}. According to this model, the reactant-like complex presents a C6(pte)-C4(nic) distance of 4.379/4.974 Å (AM1 and HF/4-31G, respectively), a C6(pte)-H-C4(nic) angle of 128.63/108.82° (AM1 and HF/4-31G) and a C8A(pte)-C6(pte)-C4(nic)-N1(nic) dihedral angle of 117.2/62.9° (AM1 and HF/4-31G). Upon formation of the transition state, the distance between the carbon atoms shortens to about 2.7 Å, the angle increases to approximately 168° and the dihedral angle between the two rings changes to about 350°.

Although these calculations were done for an \textit{in vacuo} model and neither AM1 nor HF/4-31G methods can be expected to provide very accurate results, they should be adequate to estimate the range of distances (4.5±1.5 Å), angles (>100°) and dihedrals (50±60°) that would characterise a productive ternary complex of R67•DHFH+•NADPH during MD simulations (Figure 3-18). We note that in the reported calculations good agreement was found between the same \textit{in vacuo} model, but with the \textit{exo} conformation of the rings, and a QM/MM model of the reaction within a type 1 DHFR.
Figure 3-18 MD simulations of selected complexes.
The changes in C6(pterin)-C4(nicotinamide) distance (A), C6(pterin)-H9-C4(nicotinamide) angle (B) and C8A(pterin)-C6(pterin)-C4(nicotinamide)-N1(nicotinamide) dihedral angle (C) in eight selected complexes were monitored during 4 ns simulations. Complexes adt_109 and adt_158 were originally generated using AutoDock; complexes flx_d1n2, flx_n1d2 and flx_super were obtained with FlexX; complexes man_react and man_TS correspond to manually docked structures; howH was constructed using published partial structures [507].
**C6(pte)-C4(nic) distance:**

Four out of the eight complexes (adt_158, flx_d1_n2, howH and man_TS) presented a C6(pte)-C4(nic) distance smaller than 5 Å after 4 ns of simulation (Figure 3-18A). While most structures showed a more or less stable interacting distance during the simulation, in adt_158 the distance decreased remarkably after 2.6 ns of simulation from 5.5 Å to about 3.8 Å. Therefore, all but one structure (flx_n1d2) had reasonable distances between the two reacting carbon centres after 4 ns of simulation.

**C6(pte)-H-C4(nic) angle:**

Most complexes presented angles within the expected range, between 110° and 170° (Figure 3-18B). As can be seen, the structures that presented the shortest C6(pte)-C4(nic) distances (adt_158, flx_d1n2, howH and man_TS) are the ones with the largest angles, while the structures that exhibit the smallest angles, flx_super and man_react, are among the complexes with the longest carbon-carbon distances.

**Dihedral angle C8A(pte)-C6(pte)-C4(nic)-N1(nic)**

This angle is one of the most stable parameters during the simulations (Figure 3-18C). All the complexes exhibited dihedral angles corresponding to the *endo* conformation of the interacting rings. There is a very good correlation between the C6(pte)-C4(nic) distance and the dihedral of the complexes: those structures with the shortest distances (adt_158, flx_d1n2, howH and man_TS) are the ones that presented the smallest dihedrals, closest to 0°, while the structures with the longest distances (flx_n1d2, man_react, flx_super and adt_109) had the largest dihedrals. Moreover, as the C6(pte)-C4(nic) distance of adt_158 decreases after 2.8 ps of simulation, the inter-ring dihedral angle does so as well. It seems that for a closer interaction to occur, the two rings have to adopt a specific relative orientation, where the atom centres of the two rings are almost on top of one another.

Overall, the values obtained for the interactions between the two reacting rings are within the expected range. Surprisingly, most of the complexes seemed to
have an adequate and stable disposition of the rings after 4 ns of simulation. Nevertheless, when the average structures of the last 100 ps were compared important differences were found. Although the reacting rings were located at the centre of the pore, the general conformation of the ligands, particularly that of the tails, was significantly different among the different complexes.

3.7.2.2.3 Interactions between the ligands and the protein

After analysis of the interactions between the two reacting rings, the contacts between the ligands and the protein were studied (see Table 3-3 for experimental evidence). The presence of H bonds between the ligands, DHFH+ and NADPH, and the enzyme, R67 DHFR, were monitored during the simulation (Figure 3-19). The most stable H-bond interactions were identified in all 8 different simulations and are described below.

**DHFH+ - R67 DHFR interactions:**

Residues K32, Y46, T51, S65, V66, Q67, I68 and Y69 were seen to form stable interactions with DHFH+ in at least one of the eight simulations. Those residues participating in interactions present in six or more cases included:

- K32: its interaction was present in all simulations. Its NH₃⁺ terminal group interacts with the glutamic moiety. It was a quite stable interaction during the simulations.

- V66: its interaction was present in all but one simulation (flx_super). The carbonyl group of the backbone appears to be interacting with HN5, HN8 or HN10, probably stabilizing the protonated structure and assisting in the positioning of the ring.

- I68: present in all but flx_n1d2. Interacts through its amide backbone (NH) with O4 of the pterin ring. This interaction was stable over the whole simulation period.

- Y69: present in all but two cases (flx_super and howH). The hydroxyl group of the phenyl ring interacts with the glutamic group of DHFH+; it is not as stable as the interaction of K32.
H-bond pairs between DHFH and R67 DHFR

- K32C – Glu
- I68A – Glu
- Y69C – Glu
- I68A – HN3 (pte)
- V66A – HN5 (pte)
- V66A – HN10
- G64C – ribose (nic)
- S65C – ribose
- P70A – Pi
- I68D – Ppi
- V66C – ribose
- K32A – Ppi

H-bond pairs between NADPH and R67 DHFR

- V66D – nicotinamide
- G64C – ribose (nic)
- S65C – ribose
- P70A – Pi
- I68D – Ppi
- V66C – ribose
- K32A – Ppi

Figure 3-19 Evolution of H-bond interactions along an MD trajectory.

H-bond interactions between the ligands N5-protonated dihydrofolate (A) and NADPH (B), and R67 DHFR during 4 ns of simulation of the complex adt_109. The presence of an H-bond interaction at any time is represented by a black mark. The most stable and long lasting interactions are specified by the number and chain of the residue involved and the group or atom of the ligand participating in the interaction. pABA; para-amino-benzoyl tail of DHFH+, pte: pterin ring, Glu: glutamic moiety; ribose (aden): ribose moiety bound to the adenine ring; ribose (nic): ribose moiety bound to the nicotinamide ring; Pi: phosphate group; PPi: pyrophosphate group.

NADPH - R67 DHFR interactions:

The cofactor NADPH was found to have H-bond interactions with several residues among the 8 complexes, including K32, A36, G64, S65, V66, Q67, I68, Y69, P70, A72, and A73. Those interactions present in at least five of the simulations included:

- K32: interacts with NADPH in all simulations. Its NH3+ group is interacting with the pyrophosphate group (PPi), the phosphate group (Pi) or both. The
interaction is present most of the time, except for adt_158, where the time at which the interaction appears matches the closer approach of the rings.

- V66: appears in all but one simulation (flx_d1n2). In most cases its backbone forms an H bond with the ribose subunit bound to the nicotinamide ring or with the amide group of the nicotinamide ring. This is a stable interaction.

- Q67: present in 5 of the 8 cases, except adt_158, flx_d1n2 and flx_n1d2. The amine of this residue interacts with the adenine ring (adt_109), PPI (flx_super and man_react), and the nicotinamide ring (flx_super, howH and man_TS).

- I68: interacts with NADPH in 5 simulations, excluding flx_super, howH and man_TS. In three cases it has a very stable interaction with the amide group of the nicotinamide ring (adt_109, flx_d1n2 and man_react), while in others it also interacts with PPI (adt_158) and the adenine ring (flx_n1d2).

The interactions observed for residues K32, Q67, I68 and Y69 are in clear agreement with experimental results (Table 3-3). The H-bond contacts between NADPH and R67 DHFR were more dynamic and varied than those observed for DHFH+. The same residues were seen interacting with different regions of NADPH in different complexes, suggesting that multiple configurations of the tail of NADPH were possible. When the average structures of the complexes during the last 100 ps of simulation were compared, it was clear that the characteristics of the active site allowed for a wide distribution of structures. While the conformation of DHFH+ was similar in the different complexes, the tail of NADPH adopted various positions. It is important to note that the same amino acids (K32, Q67, V66, I68 and Y69) were involved in nearly all interactions between both DHFH+ and NADPH and the protein.

### 3.7.2.2.4 Water distribution and water-mediated interactions

The active site of R67 DHFR is a spacious hourglass pore filled with water molecules. In order to analyse the changes in water distribution upon ligand binding, we determined the number of water molecules within a radius of 5 Å of those residues lining the active-site cavity (K32, A36, Y46, G64, S65, V66, Q67, I68 and Y69) in the apo form, and in the different R67•DHFH+•NADPH
complexes (Figure 3-20A). While the average distribution of water molecules remains relatively unchanged around those residues positioned close to the openings of the pore, there is an important displacement of waters at the central region of the cavity. While the pterin and nicotinamide rings adopt a stacked conformation at the centre of the active-site pore, where there is little if any water access, the tails of the ligands extend towards opposite ends in a solvent-rich environment (Figure 3-20B).

Although there are water molecules mediating interactions between the tails of the ligands and the protein, the multiple conformations lead to non-conserved interactions among different simulations. In the case of the reacting rings, only the pterin ring was seen to form a stable water-mediated interaction involving N3 and the residues Q67 or I68.

3.7.3 Discussion

Despite catalysing the reduction of DHFH+ using NADPH as a cofactor, DfrB DHFR cannot be considered a typical enzyme. It is not the kind of reaction it promotes which makes this such an unusual enzyme, but its origins and characteristics which set it apart. Although its structure and mechanism of action have been the focus of many studies, details of the way in which it binds both the reactant and the cofactor and promotes the reaction remain unknown. The elusiveness of the system to experimental definition prompted us to perform a computational dissection of DfrB and its ligands in order to define how the enzyme accommodates the reactants within the active site.

3.7.3.1 Interligand interactions

The initial modelling of the ternary complex R67•DHFH+•NADPH using docking techniques was not a trivial process due to the size and symmetry of the active-site pore. It was not possible to position the ligands de novo within the apo enzyme, but rather it was necessary to consider the presence of a core fragment of the partner, such as the pterin ring or the NMN moiety, to produce sensible structures. The need for the presence of a second molecule within the active site for the proper positioning of the ligands correlates very well with the interligand cooperativity observed experimentally [502].
Figure 3-20 Water distribution within the active-site pore.

(A) Average water distribution along the active-site cavity for both the apo and ligand-bound enzyme. The number of water molecules within a radius of 5Å from selected active-site residues (K32, A36, Y69, G64, S65, V66, Q67, I68 and Y69) was averaged over 1 ns for the simulation of both the apo enzyme and the adt_109 complex. The order of the residues in the plot corresponds to their distribution along the pore, with those residues closer to the openings located on the left side and those at the middle of the pore on the right side. (B) Final disposition of the ligands inside R67 DHFR after 4 ns simulation of the complex adt_109. The surface representation of the protein has been clipped to make the central active-site pore visible. Water molecules within 4 Å of the ligands and the active-site pore have been included. Whereas the reacting rings occupy a central position with very little water access, the tails of the ligands are located towards the openings of the pore in a solvent-rich environment.
The interaction between the two reacting molecules, particularly that of the pterin and nicotinamide rings, seems important for adequate binding and positioning of the ligands within the active-site pore.

The 4 ns MD simulations of eight different complexes provided further insights into the behaviour and stability of the ligands within the protein environment and both ligand-ligand and ligand-protein interactions important for establishment of a reactant-like R67•DHFH++•NADPH ternary structure. It was seen that most of the structures analysed presented reasonable distances between the reacting centres after 4 ns of simulation, having the C6(pte) and C4(nic) atoms less than 6 Å apart. The structures with the shortest distances (adt_158, flx_d1n2 and howH) showed larger C4(pte)-H-C6(nic) angles and smaller inter-ring dihedrals than those of the complexes with longer C6(pte)-C4(nic) distances (flx_super, man_react and flx_n1d2). The complexes flx_super and flx_n1d2 presented not only the longest distances but also the most significant fluctuations during the MD simulations, suggesting that these two structures do not correspond to stable conformations. Thus, the MD simulations allowed us to discriminate between structures, which despite showing conformity with the empirically observed constraints (see Figure 3-15C for a comparison between adt_109, howH and flx_n1d2), present dissimilar stabilities.

Docking calculations and MD simulations are, thus, complementary, with the first providing a list of more or less reliable starting structures and the second allowing further discrimination on the bases of relative stability after relaxation of the molecular complexes. It is likely this approach will be used more in the future literature, particularly in rational drug design. This topic was further developed and published in a recent review [604] in which I collected examples in literature of different approaches that combined docking and MD simulations in drug design.

### 3.7.3.2 Protein – ligands interactions

Despite most of the structures showing very similar interactions between the two reacting rings, the general conformation of the tails of the ligands, particularly that of NADPH, were significantly different. Nonetheless, when the H-bond interactions between the ligands and the protein were studied it was seen that in
most cases the same amino acids were engaged, supporting the idea of an active
site where the same residues can accommodate multiple binding conformations
of the tails of the ligands. Moreover, the same residues in equivalent monomers
were found interacting not only with different conformations of a single ligand,
but with both DHFH+ and NADPH. This dual-binding role played by equivalent
amino acids has been previously suggested [548], and is now supported by the
results of our MD simulations.

The most important residues involved in the protein-ligand interactions include
K32, V66, Q67, I68 and Y79. The interaction of K32 and Y79 with the charged
 glutamic and phosphate groups of the pABA-Glu and 2',5'-ADP tails clearly
agree with the available experimental data (Table 3-3) offering, therefore, a
better model of the ternary complex than that previously published [507].
Residues V66, Q67 and I68 are located near the centre of the pore and interact
with the reacting rings.

Despite the importance of the direct interaction between DHFH+ and NADPH
for the overall binding process and the correct positioning of the rings, the roles
of residues V66, Q67 and I68 should not be discounted.

In the case of V66 and I68, both residues appear to interact through their
backbone atoms. Therefore, as shown experimentally [582], active mutants for
these positions are not unexpected. The peptide carbonyl group of V66 appears
H bonded to HN5 of the pterin ring, probably playing a key role in the
stabilization of the reactive protonated structure. It also establishes H-bond
interactions with the hydroxyl group of the ribonicotinamide moiety of NADPH,
helping in the correct positioning of the cofactor ring. The NH backbone group
of I68 formed one of the most stable interactions seen with O4 of the pterin ring,
which may indicate its importance in the correct positioning of the ring. In the
case of NADPH, I68 interacts through its backbone with the nicotinamide ring,
again, assisting favourable placement of the reacting group.

Residues Q67 were not observed to adopt the paired H-bonded conformation
described in the crystal structure, but instead the amide groups formed H-bond
interactions with the reacting molecules in most cases. This change in
conformation of Q67 upon ligand binding agrees with NMR analysis (Table 3-3). Interactions of the pterin ring and the nicotinamide ring with Q67 were found in five complexes; these H bonds were more varied and less stable than those observed for V66. Moreover, mutants for this residue which still present wild type-like properties have been described [582], suggesting that the role of Q67 is not entirely critical.

The only persistent water-mediated interaction among different simulations involves the N3 atom of the pterin ring and residues Q67 or I68.

3.7.3.3 Multiple conformations

Other studies have shown that a particular ligand may bind to an enzyme by adopting flexible conformations. Recent studies have suggested that the increased flexibility that many inhibitors show when bound to the active site [605-607] may be important for the design of new and more potent drugs, which are less susceptible to usual enzyme mutations producing drug resistance. Not only drugs, but also natural substrates of enzymes have been found to adopt multiple binding forms. Many enzymes exhibit quite a wide substrate specificity, and, therefore, have evolved to accommodate a broad variety of reactants [608-610].

The possibility of multiple conformations is associated with the flexibility of ligands, which can adopt different orientations to adjust and maximise the interaction with a given conformation of the active site at a low energy cost. This adaptation, and the variety of probable conformations, will depend also on the properties of the active site and its capacity to accommodate different ligand positions. The arrangement of the monomers in the ternary structure of R67 results in the presence of two symmetry-related equivalent residues within each half of the pore, either of which can establish interactions with DHFH+ or NADPH. This factor, combined with the spaciousness of the active site, can account for the possibility of multiple binding modes. This potential was clearly seen during both docking and MD simulation studies. When docking DHFH+ into R67•NMN (Figure 3-15B), the glutamic moiety was found to adopt two alternative conformations, interacting with one of two equivalent K32 groups from different monomers; this agrees with Raman studies of DHF in R67 [602].
The docking of NADPH provided a large variety of structures, none of which presented the same conformation of the 2',5'-ADP moiety. MD simulations of the complexes provided further evidence of their mobility within the active site. While the reacting rings located at the centre of the pore present deviations similar to those of the backbone, the tails of the ligands show significant fluctuations over the whole simulation time. Two structures did not show this pattern, flx_n1d2 and flx_super; both of these showed significant deviations for both reacting rings also. The flx_super complex was originally very different from the rest of the simulated structures, having the rings positioned outside the centre of the pore (Figure 3-15C); the MD simulation demonstrated this was an inadequate placement, showing it to be a useful tool to discriminate the most stable conformations from a set of docking solutions that were consistent with experimental data but non-unique.

The water distribution analysis clearly showed that upon ligand binding most of the water molecules located at the centre of the cavity are displaced by the reacting rings, while the average water distribution along the rest of the active site remains more or less constant. These changes leave the negatively charged tails within a solvent-rich environment while the reacting rings are positioned in a region with little if any water access.

The positioning and behaviour of the ligands within R67 DHFR is, therefore, very different from that observed in the Type 1 or chromosomal enzyme, which presents a narrow binding groove where both molecules are positioned in different specific locations [516]. The residues lining these binding sites appear to have been optimised to accommodate each ligand selectively; therefore, the mobility of their tails is substantially hindered compared with that observed within R67 DHFR.

Our results are in good agreement with the available experimental information (Table 3-3). The mobility of the tail regions has been observed not only in both X-ray crystallography and NMR studies but is also in agreement with mutational analysis. Asymmetric mutations of Y69 indicate that variable positions of the pABA-Glu tail of DHF are tolerated [586], and studies of K32 showed that two mutations on different half pores produce topologies that allow comparable
interactions between K32 and the Glu tail of DHF [546]. Furthermore, mutations at the centre of the pore (Q67H) are accompanied by inhibition of the catalytic activity due to non-productive binding [580], while mutations close to the pore surface (K32M and Y69F) are not.

3.7.3.4 Role of R67 DHFR in the catalysis of the reaction

The size and shape of the active site, the position of charged residues, the lack of a direct proton donor or other residue directly involved in the catalytic process, and the multiple interactions that can be established between the protein and ligands support the idea of this being a recently evolved enzyme. Despite DfrB DHFR lacking many of the characteristics of more efficient catalysts, it is an enzyme that uses a simple yet effective approach to achieve catalysis of the reaction. The double-funnel shape of the active site combined with flexible H-bond interactions between the protein and the ligands help to bring the reactants together and keep them within the pore, thus promoting the encounter of the reacting rings at the centre. V66, Q67 and I68 may contribute to the formation of the reactive complex by stabilizing and confining the stacked rings in the middle of the pore, where they adopt an endo-like conformation in an environment with little if any water access, while the charged pABA-Glu and 2’,5’-ADP tails extend towards the opening of the pore adopting multiple conformations in a solvent-rich environment, where K32 and Y69 play important roles (Figure 3-20B).

3.7.4 Conclusions

Our docking analyses and MD simulations clearly suggest that there is more than one possible conformation of the ligands DHFH+ and NADPH within the active-site pore of R67 DHFH that agrees with experimentally determined properties. The feasibility of these different complexes was further supported by MD simulations, which provided evidence of their stability in spite of the different positioning of the ligand tails. While the reacting rings adopt a stacked endo conformation in the vicinity of the centre of the active site, assisted by residues V66, Q67 and I68, the tails extend towards opposite ends of the pore adopting multiple conformations, which are stabilized by H-bond interactions with K32 and Y69 in a solvent-rich environment.
The lack of specific binding sites for each ligand, the importance of interligand cooperativity, and the flexibility of the molecules within the active-site pore suggest that the main role of the protein is to facilitate the approach of the ligands. This is achieved by using what seem to be simple geometrical constraints combined with specific hydrophilic and/or hydrophobic interactions. The hourglass shape of the pore promotes the encounter of the molecules by facilitating the stacking of the reacting rings at the narrow centre, whereas the charged tails remain within the hydrophilic environment of the openings.

### 3.8 Final considerations

We have presented a comprehensive analysis of the sequence and structural properties of the DfrB family of enzymes, a type 2 DHFR that confers resistance to the widely used antibacterial drug trimethoprim. Their properties clearly do not conform to the set of specialized properties which typify the group of catalysts conventionally identified as "enzymes".

In terms of catalysis, DfrB is 1000 less efficient than chromosomal DHFR (Table 3-2) [469,497]. It probably promotes the reaction by facilitating the approach of the ligands within a cavity with an altered dielectric constant compared with solution and/or which might restrict the accessible conformations of the reactants, rather than by using specific catalytic amino acids with defined roles to promote reaction. Active mutants for all the residues that have been suggested to play an important role in reaction have been found [548,582,583]. Moreover, the correct positioning of the ligands seems highly dependent on their interligand interactions. Docking and MD simulations [243] indicate that there is more than one stable conformation of the ligands within the active site pore, suggesting that the main role of the enzyme is to facilitate the approach of the reactants for a long enough time for the reaction to occur.

In terms of structure, the active enzyme is a homotetramer, a toroidal structure with a single central active-site pore where both ligands bind to symmetrically equivalent positions. Each monomer presents an SH3 domain, a fold found mainly as an auxiliary domain in eukaryotic proteins, and never before as the sole domain of a protein, let alone as a catalytic one. It has been shown that the N-terminal region of the protein is not necessary for enzymatic activity, but
plays some role in protein stability inside the cell [503,504,583,611]. Ignoring the first 18 amino acids, 82% of the residues tolerate mutations, up to 17% of the positions at the same time [583].

From a genetic point of view, the \textit{dfrB} gene has been found only within integrons, in gene cassettes. The level of sequence divergence among DfrB proteins and the unique properties of cassettes suggest that this enzyme existed well before the clinical introduction of trimethoprim. Therefore, its original function or functions were probably different from its currently identified activity, the reduction of dihydrofolate. We could speculate that this protein class has fulfilled other useful functions in bacteria subject to different environmental challenges in the distant past, or indeed could have other catalytic activities in current bacterial environments.

There are still several questions that remain to be answered in order to fully understand the DfrB DHFR family of proteins, including:

- Apart from DfrB, are there other examples of “minimalist” enzymes which might be found within bacterial cassettes?
- Where was the \textit{dfrB} gene originally sequestered from, and what was its function?
- Are there other functional enzymes that possess an SH3 catalytic domain?
- Do gene cassettes within bacterial communities constitute a major untapped public source of proteins with novel functions – for good or bad?
- If so, what impact might such genes have on the further rise and spread of antibiotic resistance? Is there a way to control this, e.g. such as inhibition of the integrases?
- What other reactions might the DfrB enzymes catalyse?

I suspect that DfrB enzymes could be the “tip of the iceberg” and that mobile cassettes and associated genes might offer a vast adaptive resource to microbial communities. Characterization of other proteins encoded within cassettes in similar detail might provide further fruitful "minimalist" examples of biological catalysts, proteins that have been recently selected from a "communal" bacterial
gene reservoir to perform new functions. This expanded knowledge together with further studies on DfrB DHFR would not only increase our understanding of enzymatic catalysis, but would help in efforts to prevent, or at least delay, the emergence and rapid spread of antibiotic resistance through greater understanding of the mechanisms and tools that bacteria use to adapt and survive new environmental conditions.