

Université de Montréal

**Tirer profit de l'espace de séquence : une approche multidisciplinaire pour
élucider l'évolution d'une famille d'enzymes primitives**

Par

Claudèle Lemay-St-Denis

Département de biochimie et médecine moléculaire

Faculté de Médecine

Thèse présentée en vue de l'obtention du grade de *Philosophiae Doctor* en biochimie

Janvier 2024

© Claudèle Lemay-St-Denis, 2024

Université de Montréal

Département de biochimie et médecine moléculaire

Faculté de Médecine

Cette thèse intitulée

**Tirer profit de l'espace de séquence : une approche multidisciplinaire pour élucider
l'évolution d'une famille d'enzymes primitives**

Présentée par

Claudèle Lemay-St-Denis

A été évaluée par un jury composé des personnes suivantes

James Omichinski

Président-rapporteur

Joelle Pelletier

Directrice de recherche

Charles Dozois

Membre du jury

Belinda Chang

Examinatrice externe

Résumé

L'habileté des enzymes à évoluer joue un rôle fondamental dans l'adaptation des organismes à leur environnement, leur permettant de s'adapter aux changements de température, aux nutriments disponibles ou encore à l'introduction de composés cytotoxiques. Au cours des dernières décennies, cette capacité a conduit à l'émergence rapide de mécanismes de résistance aux antibiotiques chez des bactéries pathogènes pour l'humain, notamment dans le cas de l'antibiotique synthétique triméthoprim. Dix ans après l'introduction de cet antibiotique, l'enzyme dihydrofolate réductase de type B (DfrB) a été identifiée comme conférant une résistance aux bactéries l'exprimant en catalysant par voie d'enzyme alternative la réaction inhibée par l'antibiotique.

Des études structurales, cinétiques et mécanistiques de la DfrB en ont révélé la nature atypique, et suggèrent que cette enzyme est un modèle d'enzyme primitive. En particulier, son site actif unique est formé via l'interface de quatre protomères identiques. Puisque les DfrB ne sont pas apparentées sur le plan évolutif à des protéines connues et caractérisées, on ne connaît pas comment elles ont évolué pour ultimement contribuer à la résistance au triméthoprim, et en particulier comment leur capacité catalytique a émergé au sein du petit domaine codé par leurs gènes. Ainsi, cette thèse vise à approfondir notre compréhension de l'évolution des enzymes en examinant spécifiquement l'évolution des DfrB et les propriétés qui ont guidé ce processus.

Puisque les gènes des DfrB ont rarement été rapportés, je présente d'abord nos efforts déployés pour identifier et caractériser de manière génomique les DfrB dans les bases de données publiques. Ces efforts ont conduit à la découverte, pour la première fois, de DfrB en dehors du contexte clinique. Nous avons ensuite caractérisé, sur le plan biophysique et enzymatique, des homologues protéiques aux DfrB que nous avons identifiés dans des bases de données de protéines putatives. Nous avons démontré la capacité d'homologues identifiés dans des contextes environnementaux, non associés aux activités humaines, à catalyser la réduction du dihydrofolate de la même façon que les DfrB. Enfin, une large exploration d'homologues de séquence, suivie d'une caractérisation expérimentale et computationnelle, nous a permis d'identifier des homologues distants des DfrB, certains capables de procurer une résistance au triméthoprim, et d'autres dépourvus de cette capacité. Ces résultats nous ont permis de proposer un modèle expliquant l'émergence de l'activité catalytique au sein du domaine protéique des DfrB.

En résumé, cette thèse présente une approche multidisciplinaire pour l'exploration et la caractérisation de l'espace de séquence d'une famille de protéines. Cette approche, qui comprend des analyses génomiques, enzymologiques, biophysiques et bio-informatiques, nous a permis d'identifier les caractéristiques structurales et de séquences nécessaires à la formation d'une enzyme DfrB fonctionnelle. Nous avons

également proposé un modèle pour expliquer l'évolution de cette enzyme primitive. Dans l'ensemble, nos résultats suggèrent que la capacité catalytique des DfrB a évolué indépendamment de l'introduction de l'antibiotique triméthoprine, et donc que ce mécanisme de résistance existait dans l'environnement préalablement à son recrutement génomique dans un contexte clinique.

Ces travaux contribuent à notre compréhension fondamentale des mécanismes sous-jacents à l'émergence de l'activité catalytique au sein d'un domaine protéique non catalytique, et informent les études des mécanismes développés par les bactéries pour proliférer en présence d'antibiotiques.

Mots-clés : évolution des enzymes, espace de séquences, prédiction de structure, dihydrofolate réductase, cinétique enzymatique, biophysique, bio-informatique, test d'activité *in vitro*, test d'activité *in vivo*, résistance aux antibiotiques

Abstract

The ability of enzymes to evolve plays a fundamental role in the adaptation of organisms to their environment, allowing them to adjust to changes in temperature, available nutrients, or the introduction of cytotoxic compounds. In recent decades, this ability has led to the rapid emergence of antibiotic resistance mechanisms in human pathogenic bacteria, particularly in the case of the synthetic antibiotic trimethoprim. Ten years after the introduction of this antibiotic, the type B dihydrofolate reductase (DfrB) was identified as conferring resistance to bacteria expressing it by providing an alternative enzyme to catalyze the reaction inhibited by the antibiotic.

Structural, kinetic, and mechanistic studies of DfrB have revealed its atypical nature and suggest that this enzyme is a model of a primitive enzyme. In particular, its unique active site is formed by the interface of four identical protomers. Since DfrB enzymes are not evolutionarily related to any known and characterized proteins, it is not known how they evolved to ultimately contribute to trimethoprim resistance, and in particular how their catalytic ability arose within the small domain encoded by their genes. Thus, this thesis aims to deepen our understanding of enzyme evolution by specifically examining the evolution of DfrB and the properties that guided this process.

Since DfrB genes have rarely been reported, I first present our efforts to genomically identify and characterize DfrB in public databases. These efforts led to the first discovery of DfrB genes outside the clinical context. We then biophysically and enzymatically characterized protein homologues of the DfrB we identified in putative protein databases. We demonstrated the ability of homologues identified in environmental contexts unrelated to human activities to catalyze dihydrofolate reduction in the same manner as DfrB. Finally, a broad search for sequence homologues, followed by experimental and computational characterization, allowed us to identify distant DfrB homologues, some capable of conferring resistance to trimethoprim and others lacking this ability. These results have allowed us to propose a model that explains the emergence of catalytic activity within the DfrB domain.

In summary, this thesis presents a multidisciplinary approach to explore and characterize the sequence space of a protein family. This approach, which includes genomic, enzymatic, biophysical and bioinformatic analyses, has enabled us to identify the structural and sequence features necessary for the formation of a functional DfrB enzyme. We have also proposed a model to explain the evolution of this primitive enzyme. Overall, our results suggest that the catalytic capacity of DfrB evolved independently of the introduction of the antibiotic trimethoprim, and thus that this resistance mechanism existed in the environment prior to its genomic recruitment in a clinical context.

This work contributes to our fundamental understanding of the mechanisms underlying the emergence of catalytic activity within a non-catalytic protein domain, and informs studies of the mechanisms developed by bacteria to proliferate in the presence of antibiotics.

Keywords: enzyme evolution, sequence space, structure prediction, dihydrofolate reductase, enzyme kinetics, biophysics, bio-informatics, *in vitro* activity assays, *in vivo* activity assays, antibiotic resistance

Table des matières

Résumé	iii
Abstract.....	v
Table des matières	vii
Liste des figures	xiv
Liste des figures supplémentaires.....	xvi
Liste des figures en Annexe.....	xviii
Liste des schémas	xix
Liste des tableaux	xx
Liste des tableaux supplémentaires	xxi
Liste des abréviations	xxii
Remerciements	xxiv
Chapitre 1. Introduction.....	1
1.1 Vue d'ensemble	1
1.2 L'évolution des enzymes	1
1.3 Étude de l'évolution d'enzymes au 20 ^e siècle	3
1.4 Avancées technologiques des dernières décennies.....	4
1.4.1 Technologies de séquençage.....	5
1.4.2 La prédiction de gènes	7
1.4.3 Le développement de bases de données	8
1.4.4 Outils bio-informatiques pour prédire la fonction de protéines.....	10
1.4.5 Outils bio-informatiques pour identifier des relations évolutives entre protéines.....	11
1.4.6 Prédiction de structures et complexes protéiques.....	15
1.5 Utilisation de la technologie pour l'étude de l'évolution catalytique.....	17
1.5.1 Le modèle de la dihydrofolate réduction de type B.....	19
1.5.2 Efforts pour établir l'origine des DfrB	20

1.6 Le prochain défi technologique	22
1.7 Références.....	22
Chapitre 2. Revue de littérature sur la famille des DfrB	32
Préface	32
Article de revue 1. From a binding module to essential catalytic activity: how nature stumbled on a good thing	33
2.1 Abstract.....	34
2.2 Introduction.....	34
2.3 The first step to catalysis: creation of a binding scaffold	37
2.4 From scaffold to catalyst: positioning the reagent proximal to a reducing agent.....	39
2.5 The proximity-based catalytic mechanism of DfrB.....	43
2.6 A permissive active site: catalysis is maintained with modifications to either the active site of the cofactor	44
2.7 Learning more about binding by inhibiting: discovery of DfrB inhibitors	45
2.8 Poorly evolved catalysts: the DfrB family exhibit characteristics of primitive enzymes.....	47
2.8.1 The DfrB enzymes are not specifically evolved for efficient dihydrofolate reduction	47
2.8.2 A highly symmetrical pore is a poor design for an active site.....	48
2.8.3 The DfrB family lacks the specialized properties that are typical of modern enzymes	48
2.9 When catalysis means survival: the emergence of a powerful antibiotic resistance mechanism.....	49
2.10 Conclusions and perspectives	51
2.11 Conflicts of interest.....	52
2.12 Acknowledgements.....	52
2.13 References.....	52
Chapitre 3. Exploration génomique des gènes <i>dfrB</i> cliniques	62
Préface	62
Article de recherche 1. The bacterial genomic context of highly trimethoprim-resistant DfrB dihydrofolate reductases highlights an emerging threat to public health	63

3.1 Abstract.....	64
3.2 Introduction.....	64
3.3 Results.....	66
3.3.1 Expansion of the DfrB Family.....	66
3.3.2 Identification of Bacterial Sequences	66
3.3.3 Analysis of the Genomic Context.....	68
3.4 Discussion.....	72
3.5 Materials and Methods	74
3.5.1 Identification of putative type B dihydrofolate reductases.....	74
3.5.2 Subcloning of <i>dfrb10</i> and <i>dfrb11</i> genes.....	74
3.5.3 Minimal Inhibitory Concentration (MIC).....	75
3.5.4 Download of genomes	75
3.5.5 Protein database constructions.....	75
3.5.6 Annotation	75
3.5.7 Classification of sequences as chromosomal or plasmidic.....	76
3.5.8 Identification of pathogenic hosts.....	76
3.5.9 Phylogenetic tree	76
3.6 Author Contributions.....	76
3.7 Funding.....	76
3.8 Acknowledgments	76
3.9 Conflict of Interest.....	76
3.10 References.....	76
3.11 Supplementary material	82
3.12 Supplementary references.....	83
Chapitre 4. Caractérisation enzymatique et biophysique d'homologues environnementaux des DfrB	84
Préface	84

Article de recherche 2. A conserved SH3-like fold in diverse putative proteins tetramerises into an oxidoreductase providing an antimicrobial resistance phenotype	86
4.1 Abstract.....	87
4.2 Introduction.....	87
4.3 Results.....	89
4.3.1 Identification of distant homologues of the DfrB enzymes.....	90
4.3.2 Investigation of the DfrB-like phenotype in the DfrB-H.....	92
4.3.3 Extracting the homologous SH3-like segments from the DfrB-H.....	93
4.3.4 Kinetics and inhibition of the DfrB-H5 distant homologue	94
4.3.5 The DfrB-like domain promotes tetramerization	95
4.3.6 Biophysical properties of the SH3-like fold are similar in DfrB1 and DfrB-H5-Seg	98
4.4 Discussion.....	100
4.5 Methods	102
4.5.1 Identification of the homologues	102
4.5.2 Cloning	102
4.5.3 Minimal inhibitory concentration	104
4.5.4 Dihydrofolate reductase activity in lysate	104
4.5.5 Protein expression and purification	105
4.5.6 Kinetic parameters K_M and k_{cat}	106
4.5.7 Inhibition assays	106
4.5.8 Negative-stain electron microscopy sample preparation	106
4.5.9 Electron microscopy data collection and analysis	107
4.5.10 Size exclusion chromatography-multi-angle laser light scattering.....	107
4.5.11 Native mass spectrometry.....	107
4.5.12 Native gel.....	107
4.5.13 Size exclusion chromatography.....	108

4.5.14 Circular dichroism	108
4.5.15 Thermotolerance assay	108
4.6 Authors' contributions	108
4.7 Funding	109
4.8 Acknowledgement	109
4.9 Conflict of interest declaration	109
4.10 References.....	109
4.11 Supplementary material	116
4.12 Supplementary references.....	132
Chapitre 5. L'exploration de l'évolution moléculaire du domaine DfrB	133
Préface	133
Article de recherche 3. A walk through sequence space: identifying the essential features for the emergence of catalytic activity in an SH3 fold.....	135
5.1 Abstract.....	136
5.2 Introduction.....	136
5.3 Results.....	138
5.3.1 The DfrB domain is widely embedded in environmental organisms	138
5.3.2 A global view of the DfrB sequence space.....	140
5.3.3 A model of evolution of a catalytically competent active site.....	142
5.3.4 Investigating the native function of the DfrB domain	147
5.4 Discussion.....	150
5.5 Materials and Methods	154
5.5.1 Identification of homologues	154
5.5.2 Generating the sequence similarity network.....	155
5.5.3 Generating a DfrB-representative phylogenetic tree	155
5.5.4 Genomic annotation and analysis	155

5.5.5 Protein structure prediction.....	156
5.5.6 Trimethoprim resistance assay	156
5.5.7 Thermotolerance	157
5.5.8 Single curve assay.....	157
5.5.9 Protein purification	158
5.5.10 Kinetic characterization	158
5.5.11 Size exclusion chromatography.....	159
5.6 Authors' contributions	159
5.7 Funding	159
5.8 Conflicts of interest.....	159
5.9 Acknowledgments	159
5.10 References.....	159
5.11 Supporting information.....	167
5.12 Supplementary references.....	183
Chapitre 6. Discussion	184
6.1 L'étude des mécanismes à la base de l'évolution d'enzymes ouvre des horizons	184
6.2 L'évolution récente des gènes <i>dfrB</i>	184
6.2.1 L'identification des <i>dfrB</i>	184
6.2.2 La métagénomique pour étudier l'évolution des DfrB	185
6.2.3 L'évolution génomique des DfrB	186
6.3 L'évolution moléculaire du domaine DfrB.....	188
6.3.1 Fonction(s) native(s) du domaine DfrB.....	188
6.3.2 Ingénierie du domaine DfrB comme modèle d'émergence de la catalyse	189
6.3.3 Pousser l'exploration de l'espace de séquences des DfrB.....	191
6.4 Questions d'épistémologie.....	193
6.5 Perspectives	194

6.6 Références.....	195
Annexe 1. Le prochain défi technologique pour l'évolution et l'ingénierie d'enzymes	198
Préface	198
Article de revue 2. Integrating dynamics into enzyme engineering	199
A1.1 Abstract.....	200
A1.2 Why protein dynamics are relevant to engineering catalytic function	200
A1.2.1 The conformational landscape defines enzyme function.....	202
A1.3 How knowledge of protein dynamics could enrich enzyme engineering strategies.....	204
A1.3.1 Dynamics as a design tool	204
A1.4 The state-of-the-art in dynamic engineering of enzyme function.....	207
A1.5 Outlook: How can dynamic engineering become widely implemented?	210
A1.6 Funding	213
A1.7 Acknowledgments	213
A1.8 References.....	214

Liste des figures

Figure 1.1. Mécanisme d'émergence d'une nouvelle fonction catalytique à partir d'une protéine existante.	3
Figure 1.2. Ligne du temps avec des avancées technologiques centrales au développement de la biochimie évolutive.	5
Figure 1.3. Le coût du séquençage a chuté depuis le début du séquençage à grande échelle.	6
Figure 1.4. Les bases de données protéiques d'UniProt comportent plusieurs collections et versions.	9
Figure 1.5. Croissance des bases de données d'UniProt depuis 2011.	10
Figure 1.6. Plusieurs approches ont été développées pour identifier des homologues de séquences à partir d'une ou plusieurs séquences d'intérêt.	13
Figure 1.7. Plusieurs représentations peuvent être utilisées pour décrire une famille de protéines.	14
Figure 1.8. Les alignements de séquences multiples (MSA) informent sur la structure des protéines dans l'espace.	17
Figure 1.9. Structure de la DfrB1, une enzyme homotétramérique (PDB 2rk1).	19
Figure 2.1. The crystal structure of the DfrB1 enzyme (2rk1 PDB).	36
Figure 2.2. Conservation of the DfrB1 sequence.	39
Figure 2.3. Complex of the DfrB1 enzyme with the cofactor NADP ⁺ (yellow) and the pterin of DHF (dark blue) (2rk1 PDB).	40
Figure 2.4. The VQIY active-site motif and its variations in engineered DfrB1 variants.	41
Figure 2.5. Inhibitors of the unrelated dihydrofolate reductase enzymes DfrB and FoaA.	46
Figure 2.6. Timeline of the key discoveries concerning the DfrB enzymes.	50
Figure 3.1. Annotated phylogenetic tree of species harboring a <i>dfrb</i>	69
Figure 3.2.	70
Figure 3.3. Distance between <i>dfrbs</i> and genes associated with genomic mobility.	71
Figure 4.1. Overview of DfrB1 structure and key residues.	89
Figure 4.2. Distant homologues of DfrB.	91
Figure 4.3. DfrB-H proteins display a DfrB1-like phenotype.	93

Figure 4.4. Oligomerisation of DfrB-H5 analysed with 2D classification and light scattering.	96
Figure 4.5. Influence of heating on multimerisation and activity.....	99
Figure 5.1. Proteins containing the DfrB domain are endogenous to environmental organisms.	138
Figure 5.2. Tetramerization is a defining property for the emergence of a catalytic function in the DfrB enzymes.	141
Figure 5.3. The DfrB tunnel is an environment evolved for ligand binding rather than a specialized active site.....	144
Figure 5.4. Thermostability and dinucleotide binding in the evolution of the DfrB domain.	146
Figure 5.5. Insights into a putative native function of the DfrB domain.....	149
Figure 6.1 Les prochaines étapes pour pousser notre compréhension du domaine DfrB et de sa capacité catalytique.....	190
Figure 6.2. Paysage de <i>fitness</i> décrivant l'évolution de l'émergence de la catalyse au sein du domaine DfrB.	192

Liste des figures supplémentaires

Figure S4.1. Sequence alignment of representative bacterial FolA and DfrA.	117
Figure S4.2. Structure of characterized dihydrofolate reductases.	118
Figure S4.3. Sequence alignment of all characterized DfrB.	118
Figure S4.4. Catalytic mechanism for dihydrofolate reduction by DfrB1.	119
Figure S4.5. Sequence alignment of all 30 sequences identified and presented in Figure 4.2.	123
Figure S4.6. Sequence similarity (%) of DfrB-H.	124
Figure S4.7. SH3-like sequence matrix.	124
Figure S4.8. Sequence alignment of DfrB1 and the DfrB-H.	125
Figure S4.9. Representation of the generation of DfrB-H-Seg sequences from their respective DfrB-H sequence.	126
Figure S4.10. Tricine-SDS-PAGE of the heated lysate of overexpressed proteins for which dihydrofolate activity has been determined.	127
Figure S4.11. DfrB-H5 assembles into various multimeric assemblies.	128
Figure S4.12. Native nanoESI-MS and ion mobility studies demonstrate that the W38F substitution in DfrB1 nearly abolishes the tetramerisation observed in WT DfrB1.	129
Figure S4.13. Oligomerisation observed with native PAGE.	130
Figure S4.14. SEC chromatograms of WT and W236F DfrB-H5.	131
Figure S4.15. The members of the DfrB family are tolerant to heating.	131
Figure S5.1. The DfrB enzymes catalyze the hydride transfer from NADPH to DHF.	167
Figure S5.2. The guanine-cytosine (GC) content of DfrB genes identified in a clinical context is not proportional to the GC content of the organism in which they are embedded.	168
Figure S5.3. Computational pipeline for the analysis of the multimerization prediction for DfrB homologues.	169
Figure S5.4. AlphaFold-multimer predicts the DfrB1-like complexes in clusters <i>I</i> and <i>II</i> with high confidence.	170

Figure S5.5. The nine DfrB homologues predicted to form a DfrB1-like tetramer but that do not display trimethoprim resistance when expression is induced in <i>E. coli</i>	171
Figure S5.6. The surface electrostatic potential of predicted DfrB1-like tetramers for 13 of the 18 active site motifs.	172
Figure S5.7. DfrB domains across of the DfrB sequence space were characterized <i>in vitro</i>	173
Figure S5.8. Thermotolerance is not a defining property of the DfrB1-like tetramer.	174
Figure S5.9. The homotetrameric DfrB1 (pdb 2rk1) and monomeric <i>E. coli</i> Fola (pdb 4psy) enzymes both catalyze the reduction of dihydrofolate yet are structurally unrelated.	175
Figure S5.10. Genomic context of DfrB homologues.	176
Figure S5.11. Relationship between sequence length and acquisition of the tetramerization property in DfrB homologues.	177
Figure S5.12. NTP pyrophosphohydrolase domains fused to DfrB domains.	178
Figure S5.13. The DfrB domain is fused to a variety of characterized domains.	179
Figure S5.14. The DfrB-TS enzyme is bifunctional, acting as both a thymidylate synthase (TS domain) and a dihydrofolate reductase (DfrB domain).	180
Figure S5.15. Sequence logos are shown for each cluster of the DfrB sequence space.	181
Figure S5.16. Lys32 is functionally important for the DfrB domain.	182
Figure S5.17. Pipeline for the calculation of kinetic parameters using data from a single activity curve.	183

Liste des figures en Annexe

Figure A1.1. State-of-the-art and current challenges in dynamic engineering.....	201
Figure A1.2. Integrating dynamic engineering into enzyme design.....	205
Figure A1.3. Proposed workflow for dynamic engineering based on reference 79.	208
Figure A1.4. A brief history of enzyme engineering methods.	210

Liste des schémas

Scheme 2.1. Chemical reaction catalyzed by the DfrB enzymes.	35
Scheme 2.2. Ligand binding to DfrB.	42
Scheme 2.3. Dihydrofolate reductase and NADP ⁺ phosphatase reactions catalyzed by DfrB1.	44

Liste des tableaux

Table 3.1. Information and MICs on the newly identified DfrB10 and DfrB11	66
Table 3.2. Taxonomic classification of all strains identified that include at least one <i>dfrb</i>	67
Table 4.1. Kinetic parameters for the dihydrofolate reductase activity.....	94
Table 4.2. Inhibition of DfrB-H5 with structurally distinct, DfrB-specific inhibitors.	95
Table 5.1 Kinetic and oligomeric characterization of DfrB homologues.....	145

Liste des tableaux supplémentaires

Table S3.1. Protein sequence identity of the ten members of the DfrB family ^a	82
Table S3.2. DfrB sequences.....	82
Table S4.1. Taxonomic information on the DfrB-H characterized in this study.....	116
Table S4.2. MICs performed in IPTG induction broth.....	117
Table S4.3. MICs for deleterious mutations in DfrB1 and DfrB-H5.	117

Liste des abbréviations

<u>Abbréviation</u>	<u>Définition</u>
DfrB	Dihydrofolate reductase de type B
DHF	Dihydrofolate
<i>E. coli</i>	<i>Escherichia coli</i>
EM	Electron microscopy
FolA	Dihydrofolate reductase bactérienne
HMM	Hidden Markov Model
MMC	Modèles de Markov Cachés
IPTG	Isopropyl 1-thio-β-D-galactopyranoside
k_{cat}	Constante catalytique
k_{cat}/K_M	Efficacité catalytique
K_M	Constante de Michaelis-Menten
MD	Molecular dynamics
MIC	Minimal inhibitory concentration
MGE	Mobile genetic element
MS	Mass spectrometry
MSA	Multiple Sequence Alignment
NADPH	Nicotinamide adenine dinucleotide phosphate
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
SH3	Src Homology 3
TMP	Triméthoprim
THF	Tetrahydrofolate
WT	Wild type

Sans la curiosité de l'esprit, que serions-nous ? Telle est la beauté et la noblesse de la science : un désir sans fin de repousser les frontières du savoir, de traquer les secrets de la matière et de la vie sans idée préconçue des conséquences éventuelles.

– Marie Skłodowska-Curie

Remerciements

En arrivant à la fin de mon parcours, je tire le constat suivant : un village est nécessaire pour former une personne au doctorat. Je suis incroyablement reconnaissante à tous ceux et celles qui m'ont aidée à progresser – sur le plan scientifique et sur bien d'autres plans – au cours de ces dernières années, et j'aimerais souligner l'impact qu'ils ont eu sur mon cheminement.

La première personne que je veux remercier est Elaine Meunier, TGDE en biochimie. Si vous ne m'aviez pas suggéré de m'inscrire directement pour un doctorat alors que je posais ma candidature pour la maîtrise (me faisant ainsi reconsidérer mes plans académiques), je crois sérieusement que ces quatre dernières années et demie se seraient déroulées complètement différemment. Il y a parfois de petites actions qui ont des impacts énormes, et c'était certainement le cas de votre bref courriel au printemps 2019.

Si ces dernières années m'ont permis de me développer scientifiquement, humainement, et sur le plan relationnel, je le dois à toi, Joelle. Je ne peux exprimer à quel point je te suis reconnaissante de m'avoir dirigée durant les dernières années. Tu représentes la définition de ce qu'est une mentore exceptionnelle. La confiance que tu m'as démontrée dès mon début au laboratoire m'a permis de m'épanouir pleinement au sein de mon projet de recherche, dans des initiatives diverses, au sein de nombreuses collaborations, à l'international, de me découvrir de nouvelles habiletés et intérêts, et j'en passe. Ça a été un réel privilège de te côtoyer, d'apprendre de toi, d'apprendre à écrire avec toi – je me suis d'ailleurs rendu compte lors de l'écriture de la thèse que j'ai développé un « Joelle plug-in », car je t'entends commenter et me donner des indications sur mon écriture et mes figures en temps réel, c'est super pratique ! Parmi les choses que j'ai apprises de toi, on retrouve l'importance d'être passionnée par ce qu'on fait, et de le faire avec enthousiasme : je suis fière de terminer le doctorat avec encore plus d'entrain que j'avais au début (quel privilège !). J'ai aussi appris de toi que c'est possible – et nécessaire – d'être une scientifique qui respecte ses valeurs sociales, et que c'est fondamental de s'investir dans le milieu dans lequel nous évoluons. On est humain.e avant tout, et tu le démontres chaque jour brillamment. Merci pour ta générosité, merci pour le temps que tu m'as offert, merci pour les ressources auxquelles tu m'as donné accès et merci pour tout ce que tu m'as appris.

Ensuite, je veux remercier la grande famille du laboratoire Pelletier. C'est vraiment un esprit familial que j'ai ressenti au cours de mon parcours au laboratoire ; on a vécu beaucoup de choses ensemble, ce qui nous a certainement rendus plus forts. D'abord, merci à Lorea : tu m'as introduite au monde de la recherche, et tu as semé la graine de la passion de la recherche chez moi (je te revois encore t'exclamer « Géniaaal ! » devant chaque résultat), je t'en suis très reconnaissante. I am also super thankful to you, Donya, my academic big sister! It has been a real pleasure to work with you, learn from you, and run with you (textually,

but also academically!) for the first three years of my Ph.D. Thank you for integrating new traditions: how would we do without them? Ali, mon opposé en tous points, je te remercie pour ton enthousiasme, pour nos conversations pleines d'explosifs, et pour ton calme invariable qui a permis de contrebalancer mon agitation invariable (un excellent duo yin et yang, quoi !). Merci à Adem, avec qui j'aurai traversé le doctorat en même temps (on y est presque !), et à Jo, l'expert ultime des brownies véganes moelleux nécessaires à toute bonne recherche. Merci également à mes stagiaires, Sarah-Slim et Katia ; ça a été un plaisir de vous introduire au merveilleux monde des DfrB. Et merci aux autres stagiaires que j'ai côtoyées au laboratoire, avec leur alcool et desserts maisons de qualité supérieure : Léa, Arianne et Megan. Thanks to Doug, my academic big brother! You have been a great support, full of insightful advice and delicious homemade goods. You have even provided me with the template for this very thesis! Je remercie aussi Megan-Faye, qui a toujours été très réactive ; tu as secouru à de nombreuses reprises mon matériel de recherche ! Hadi, ton enthousiasme a été super apprécié. Merci de m'avoir permis de vivre mon rêve de créer un club de course, même si bref et microscopique ! Merci également à l'équipe de feu 'DfrB' ; en plus de former un beigne avec un site actif central, cette enzyme est capable de créer un sentiment d'appartenance et d'entraide chez ceux et celles qui l'étudient : une vraie « Jack of all trades », cette coquine ! D'abord, je te remercie Kiana pour le rôle précieux que tu as joué dans mon projet. Alexis et Samy, l'avenir de la DfrB et de la résistance aux antibiotiques étant en vos mains, vous avez toute ma confiance et mon encouragement. Maxime. Ça a été un réel privilège de travailler à tes côtés, de te voir t'épanouir en recherche, de te voir développer une assurance et une confiance dignes de tes efforts rigoureux et consciencieux. J'espère faire honneur à ton travail dans cette thèse, qui a été instrumental à son aboutissement. Merci pour tout. Et finalement, un merci des plus chaleureux à mon binôme, à ma partenaire au sein de l'exceptionnelle sous-équipe 'DfrB-évolution' : Stella. Je ne puis convenablement mettre en mot la reconnaissance que je te porte (les emojis n'y arrivent pas non plus, quelle horreur éhontée !). Refaire le monde chaque semaine et réfléchir à des questions d'évolution, toujours avec la DfrB au centre du monde, m'ont rempli d'une joie immense. Nos moments d'*Eureka* fréquents auront été un **high** de mon doctorat. Tu es le futur de l'étude de l'évolution de la DfrB ; je ne peux qu'en être *rassurée* et partir en paix.

Daniela, working with you has allowed me to dip my toes into enzyme engineering and to learn how to work extremely efficiently, the importance of effective communication, and most importantly, the need to set priorities that are not just academic. I am very grateful for the brief boredom I experienced a few years ago now, that led me to our ongoing collaboration.

Les collaborations académiques auxquelles j'ai participé m'ont également énormément enrichi. A special thank you to Janine, who basically started my doctoral project. This is an example of how the flap of a butterfly's wing can have a big impact on someone's life. Merci également à Nicolas Doucet. Tu m'as

ouvert les portes de ton laboratoire, avec les personnes accueillantes que sont Myriam, Hang, Quynh, Hieu et Sacha. Ça a été un vrai plaisir d'écrire la revue avec toi et de te côtoyer au comité de direction de PROTEO ; ton écoute, ta douceur, ta sagesse et ta sensibilité m'ont grandement impacté et inspiré. Hang, ton énergie sans limites me mystifie et me fascine. Thanks also to Christopher Thibodeaux for allowing me to play with native MS, and to Nuwani for your patience, help and expertise. Merci également à Christian Baron, et à Zakaria, qui a sauté dans le projet DfrB avec entrain et proactivité. Thanks also to Soichiro Tsuda and Keigo Ide for allowing us to push further our exploration of the DfrB evolution. Merci à Guillaume Lamoureux, qui, même si les fruits de notre collaboration ne sont pas présents dans cette thèse, m'a ouvert chaleureusement la porte à la dynamique moléculaire, et m'en a enseigné les bases.

Another collaboration that has greatly influenced my project and made me grow as a person is with the groups of Nir Ben-Tal and Rachel Kolodny. Nir, thank you for the warm welcome you gave me and my enzyme. My visit to your lab is the highlight of my Ph.D., and I am incredibly grateful that you offered me this opportunity with open arms. Rachel, tu ne soupçonnes pas l'impact que tu as eu sur moi. Tu es la personne la plus passionnée que je connaisse, et ta capacité à partager ton enthousiasme est tout simplement inspirante. Nos échanges ont représenté pour moi un cours accéléré sur la vie et sur la bio-informatique (dans cet ordre), et je t'en suis incroyablement reconnaissante. I would also like to thank the entire laboratory for their incredible welcome. Hila, for your reassuring presence; Gilad, for your dry sense of humor; Jaspreet, for your sweetness; Maze, for your energy; Gabi, for your incredible cooking; Kaiyu, for your contagious passion; Barak and Elon, for your calm presence. Rinat, you were like a big sister to me: thank you for your openness and sensitivity. Thank you, Kılıç, for your wisdom and gentleness. A special thanks, to my neighbor David, המורה שלי לעברית, who filled my time in the lab with banter and hummus. Thank you, Amit. You are a strange creature at times, but you are certainly full of wisdom. Thank you for being there when I needed you most and for discussing evolution with me. Last but not least: thank you, Caro. Your arrival at the lab was a divine gift for which I am incredibly grateful. I am very lucky to have met you and for your precious insights on my project. I already look to our past adventures with nostalgia: to many more! אני אוהבת את כולכם מאוד!

PROTEO a également eu un impact important sur mon doctorat. Il m'a permis de développer mes habiletés de leadership, de m'introduire aux processus sous-jacents à la direction d'un regroupement stratégique, de participer à une conférence internationale, et de rencontrer d'incroyables chercheur.euse.s. Sarah, ça a été un bonheur d'apprendre à être représentante étudiante avec toi ! Merci à Steve Bourgault et aux membres du comité de direction pour leur accueil. PROTEO m'a aussi permis de rencontrer le laboratoire de Christian Landry, de partager sur l'incroyable DfrB, et de rencontrer Isabelle et Angel. Thank you, Angel,

for the smart conversations on evolution, and for your witty sense of humor that we shared on three continents (with the legendary « אני שותה בירה כשאני עצוב » and many more).

Merci également à APRENTICE – et ce faisant à Roberto Chica – qui m’a permis de me développer en tant que chercheuse. Je suis très reconnaissante des opportunités que ce programme m’a offertes. Surtout, ça m’a permis de faire la rencontre de Rojo, ma partenaire académique et ma confidente, avec qui j’ai appris à lancer des initiatives de toute sorte. Je suis hyper chanceuse de t’avoir, et d’avoir parcouru le doctorat avec toi, même si à 200 km de distance.

Je remercie également le support indispensable du CRSNG, du FRQNT, de MITACS, du CCVC, d’Hydro-Québec, d’APRENTICE, de PROTEO et de l’Université de Montréal, qui m’ont permis de faire ma recherche sans stress financier, de la présenter dans quatre pays, et de faire un stage à l’international.

Merci à Marie Pageau ; je vous suis reconnaissante pour votre douceur et la confiance que vous m’avez portées, toujours avec respect et bienveillance. Merci également à Pascale Legault et Jim Omichinski, avec qui j’ai eu le plaisir de travailler dans le cours BCM2501 au cours des quatre dernières années.

Merci aux membres de comité de thèse, John Pascal et Charles Dozois, ainsi qu’au jury de mon examen prédoctoral, Christian Baron, Christian Landry et Liz Meiering. Vous m’avez permis de réfléchir à mon projet avec différents points de vue, ce qui m’a été très bénéfique.

Je veux aussi remercier mes précieux relecteurs de thèse – Stella, Hang et Joelle – qui m’ont permis d’améliorer et de préciser mes idées.

Merci au café Nocturne (feu café Perko), qui a été mon lieu culte pour l’écriture de la thèse et de mes articles. Un merci chaleureux au studio Surya Montréal, en particulier à Bernhard et à Claudette, qui m’ont permis de faire de la méditation une partie intégrante et essentielle de ma dernière année au doctorat.

Je dois remercier mes amis, qui m’ont sorti la tête de la thèse. Merci à Janie, Véronique et Ariel, Anne-Sophie, Val, Juliane V., Ophélie, Charlotte, Lou et Emma, Mélanie et Jocelyne, Bellastrid, Hang, Rojo, Stella, Samy, Caro, Rinat, Kılıç, Amit et Angel, Donya and Roza, Juliane F. and Jaime, et tous les autres qui ont rendu mon expérience mémorable.

Merci au soutien de mes familles Lemay et St-Denis, qui m’ont toujours témoigné une grande fierté.

Finalement, je dois l’aboutissement de mon parcours académique à mes parents Carole et Michel (et à nos félins). Vous avez été d’un support immuable et indispensable tout le long de ma scolarité, de la prématernelle à aujourd’hui. Je vous aime infiniment.

Chapitre 1. Introduction

1.1 Vue d'ensemble

Les processus sous-jacents à l'évolution de nouvelles fonctions catalytiques se déploient depuis des milliards d'années, mais demeurent, à ce jour, largement méconnus, en grande partie en raison de la complexité inhérente à la rétrospective des événements évolutifs anciens. Les enzymes qui ont évolué en réponse à des composés anthropogéniques constituent d'excellents modèles d'étude pour les biochimistes évolutifs, car cette évolution est, par nature, récente et observable. L'objectif principal de cette thèse est d'utiliser les technologies et avancées récentes pour percer le mystère de l'origine évolutive d'un système enzymatique identifié suite à l'introduction d'un composé antibiotique synthétique, non inspiré de la nature. L'introduction revisite les avancées technologiques majeures dans les domaines portant sur l'étude de l'évolution naturelle d'enzymes. J'y discute des innovations que ces avancées ont permises, ainsi que de leurs limites actuelles. Je présente ensuite le système enzymatique à l'étude au Chapitre 2, suivi des efforts entrepris pour définir ce qui a mené à l'émergence de ce système enzymatique aux Chapitres 3, 4 et 5. Je discute des questions actuelles concernant l'évolution de la famille des DfrB au Chapitre 6. En Annexe 1, je présente le défi technologique émergent que représente l'intégration des informations relatives à la dynamique des protéines à notre compréhension des enzymes et de leur évolution.

1.2 L'évolution des enzymes

Les enzymes, apparues il y a quatre milliards d'années, sont essentielles à la vie telle qu'on la connaît.¹ Au cours de l'évolution, elles ont dû s'adapter aux conditions dans lesquelles elles se trouvaient. La nature des raisons de leur adaptation est diverse : des changements de température, de concentration des substrats, d'identité et de concentration de sels dans l'environnement ou encore l'introduction d'un composé cytotoxique influencent leur évolution. Les environnements de survie ardu, tel que la mer morte ou les volcans, sont des lieux extrêmes où des adaptations moléculaires spectaculaires sont révélées possibles.^{2,3}

Plusieurs propriétés sont utilisées pour décrire les enzymes dans un contexte évolutif. D'abord, la spécificité des enzymes réfère à leur capacité de transformer un substrat spécifique du fait de leur affinité particulière avec ce dernier. On peut la quantifier avec la constante k_{cat}/K_M , aussi connue sous le terme d'efficacité catalytique, qui est le ratio entre la constante catalytique et la constante de Michaelis-Menten.⁴ En revanche, la promiscuité d'une enzyme décrit la capacité d'une enzyme à catalyser une ou plusieurs réactions chimiques différentes de son activité native, pour lesquelles elle n'a pas évolué.⁵ Leur évolutivité (pour «

evolvability ») concerne leur capacité à acquérir une nouvelle fonction suite à peu de substitutions.⁶ Les protéines multifonctionnelles, codées par une seule chaîne polypeptidique, exhibent plus d'une fonction catalytique ou de liaison.⁷ Les protéines « *moonlighting* » sont un type de protéines multifonctionnelles qui excluent à la fois les protéines résultant de fusion de gènes et celles ayant une fonction pouvant opérer dans différents compartiments ou accepter différents substrats.^{8,9} Dans le cadre de cette thèse, un domaine fait référence à un segment protéique qui partage une similarité de séquence significative avec d'autres séquences.¹⁰⁻¹² Un repliement protéique (ou « *fold* ») détient plutôt une définition structurale : il correspond à une unité protéique compacte formée d'éléments de structure secondaire qui se replie indépendamment.^{13,14} Le concept d'espace de séquences fait référence à l'ensemble théorique de toutes les séquences protéiques possibles.^{15,16} On peut représenter de diverses manières comment cet espace a été exploré naturellement par l'évolution ; les réseaux de similarité de séquences sont d'ailleurs couramment utilisés^{17,18}, comme présenté au Chapitre 5.

Un mécanisme majeur pour l'évolution d'une nouvelle enzyme est la duplication d'enzymes promiscuitaires, suivie de leur divergence pour perfectionner une activité ou une propriété (Figure 1.1A).¹⁹⁻²¹ La spécialisation catalytique des enzymes promiscuitaires découle de modifications structurales, allostériques ou encore dynamique.^{22,23} Ainsi, la spécificité d'une enzyme peut être modifiée, soit catalysant la même transformation chimique sur un substrat différent²⁴, ou même modifier la transformation chimique qu'elle peut catalyser²⁵.

Au contraire de la duplication d'enzymes existantes, l'émergence d'une enzyme à partir d'une protéine non catalytique est peu répertoriée.^{26,27} L'évolution d'enzymes à partir de protéines de liaison non catalytiques a toutefois été établie.²⁸⁻³⁰ L'émergence de l'activité catalytique a été décrite par l'inclusion successive de mutations « fondatrices » dans le site de liaison évoluant en site actif, suivie de la substitution de résidus distaux pour améliorer l'activité catalytique croissante.^{31,32}

Dans les deux cas, il s'agit d'une fonction catalytique qui est gagnée à partir d'une protéine de départ ayant une faible activité promiscuitaire, et dont l'activité catalytique peut être augmentée si elle est bénéfique à l'organisme (Figure 1.1).

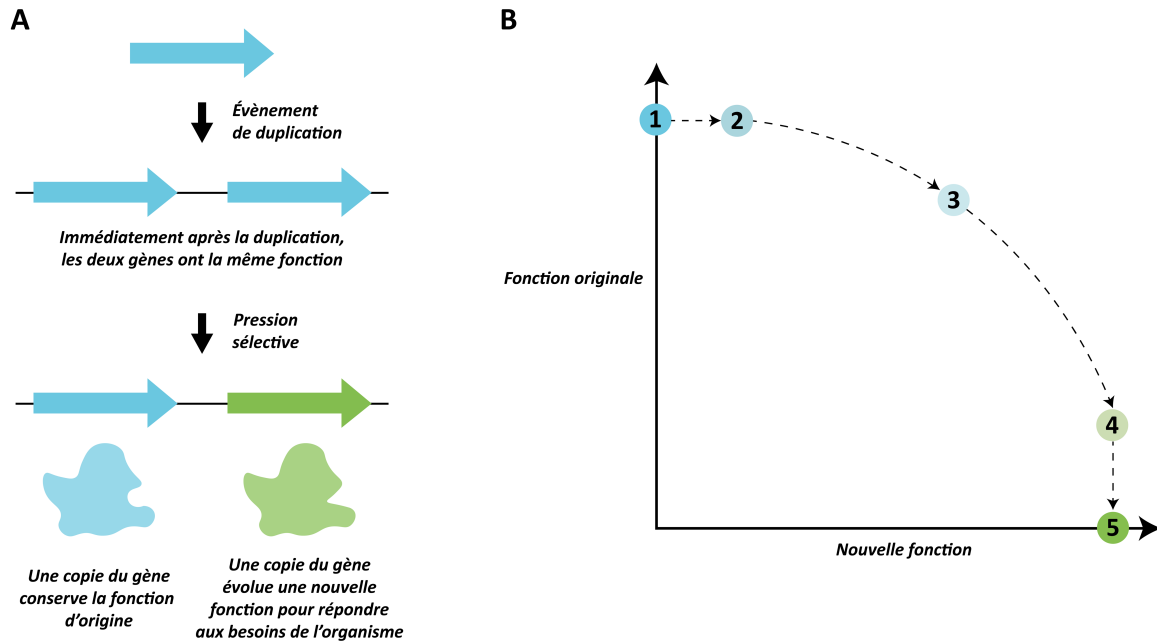


Figure 1.1. Mécanisme d'émergence d'une nouvelle fonction catalytique à partir d'une protéine existante.
A. Immédiatement suite à la duplication d'un gène, les deux copies ont la même fonction. Avec le temps et une pression sélective, des mutations s'accumulent. Une des deux copies conserve la fonction originale, nécessaire à l'organisme, alors que l'autre copie, ne subissant pas la pression évolutive de conserver l'activité originale, peut évoluer une nouvelle fonction bénéfique. **B.** Une représentation simplifiée de la trajectoire évolutive vers une nouvelle fonction, de l'état initial (1) à l'état final (5), inspirée de ²⁶. La dérive neutre, soit l'accumulation de mutations aléatoires qui ne sont pas dirigées vers une fonction particulière, permet à la protéine de départ de gagner une faible activité promiscuiteuse (2), sans impact sur son activité originale. Puisque cette nouvelle activité est bénéfique, une accumulation de mutations permet d'en augmenter son efficacité, au détriment de l'activité de départ. La protéine passe par un état d'intermédiaire généraliste (3), et à force d'accumuler des mutations pour augmenter son efficacité pour la nouvelle fonction, compromet sa capacité à effectuer la fonction originale (4), qui sera éventuellement perdue (5).

1.3 Étude de l'évolution d'enzymes au 20^e siècle

Au 20^e siècle, les études sur l'évolution d'enzymes utilisaient des approches quasi exclusives à la biochimie. Ainsi, pour quelques membres d'une famille d'enzymes, on déterminait expérimentalement, entre autres, leur structure, leur activité cinétique, leur stabilité, leur spécificité pour plusieurs substrats et leur pH optimal.^{33,34} Les différences entre quelques enzymes étaient alors soulignées et rationalisées selon l'environnement dans lequel elles ont évolué. Cette approche est exigeante en ressources et en temps, expliquant pourquoi un nombre limité de systèmes enzymatiques a été étudié de telle façon. Parmi les systèmes enzymatiques dont l'évolution a été étudiée, on retrouve les ribonucléases³⁵, les β -lactamases³⁶ et les chymotrypsines³⁷.

Par exemple, la dihydrofolate réductase monomérique (DHFR) est l'un des systèmes les mieux étudiés ; la DHFR d'*E. coli* a été extensivement comparée à celle des humains ou d'autres vertébrés.³⁸⁻⁴¹ Ces études ont permis de mettre en lumière les différences relatives à la liaisons des substrats, les différences structurales, ainsi que les différences importantes relatives à la dynamique protéique entre l'enzyme d'*E. coli*, dont la flexibilité est centrale à la catalyse, et les enzymes des vertébrés, plus rigides.⁴² Les différences observées dans la flexibilité et les populations de conformations ouverte et fermée chez les DHFR de différents organismes sont maintenant expliquées par des différences en concentration des produits NADP⁺ et tétrahydrofolate, qui sont plus élevées chez *E. coli*, nécessitant la favorisation de la conformation fermée chez la DHFR de cet organisme.⁴³

Quant à elles, les kinases d'adénylate ont été particulièrement étudiées pour leurs adaptations moléculaires entre membres provenant d'extrêmophiles, tel que des archées thermophiles.^{44,45} Les études structurales ont mis en lumière l'importance d'un cœur hydrophobe compact chez les kinases d'adénylate de thermophiles ; ces études ont suggéré que l'augmentation de résidus avec chaînes latérales aliphatiques est un contributeur majeur de la thermostabilité de ces enzymes.^{46,47}

Bien que les approches et techniques disponibles au 20^e siècle aient permis d'étudier l'évolution d'un éventail limité d'enzymes, cette période a permis d'établir les bases du domaine de la biochimie évolutive. Tout d'abord, il a été établi que la structure d'une protéine est déterminée par sa séquence en acides aminés.^{48,49} Ensuite, la comparaison de structures de protéines homologues a mené à l'observation que la structure tridimensionnelle des protéines est plus conservée que leur séquence.⁵⁰ Également, il a été découvert qu'à la base de l'évolution de protéines se trouve la duplication de gènes (Figure 1.1).^{51,52} Un élément important de cette duplication est la promiscuité, qui, opposée à la spécificité, est nécessaire pour permettre à une enzyme de catalyser une réaction chimique différente de la réaction originale, même si l'activité catalytique est faible. Pour améliorer la capacité d'une enzyme à faire une transformation chimique, celle-ci doit être capable de faire cette transformation, même si faiblement. D'ailleurs, cette plasticité est centrale à la capacité des organismes et enzymes à s'adapter aux pressions de leur environnement.

1.4 Avancées technologiques des dernières décennies

Les dernières décennies ont été ponctuées par des avancées technologiques révolutionnant la science à divers niveaux (Figure 1.2). Ces innovations ont créé un effet synergique permettant une compréhension du monde qui nous entoure par l'étude des protéines qui ont été naturellement explorées au cours de l'évolution. Ci-dessous, je décris de manière non exhaustive certains développements majeurs qui ont transformé notre façon d'étudier les protéines naturellement évoluées.

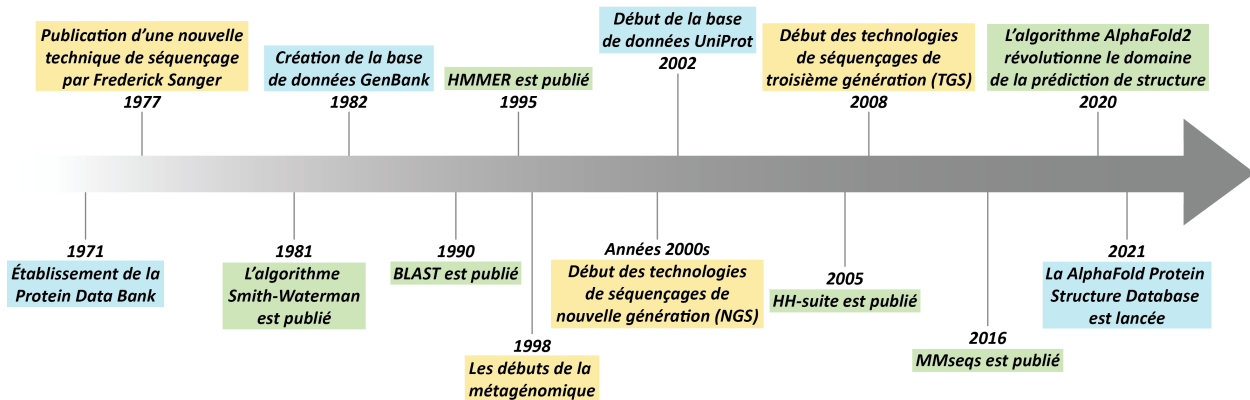


Figure 1.2. Ligne du temps avec des avancées technologiques centrales au développement de la biochimie évolutive.

Les avancées reliées aux bases de données sont indiquées en bleu, celles reliées au séquençage en jaune, et celles reliées aux algorithmes en vert.

1.4.1 Technologies de séquençage

Une révolution technologique majeure en génomique, et incidemment en biologie moléculaire, a été notre capacité à séquencer des génomes de manière rapide et peu coûteuse. Cette révolution a été initiée par la publication de la méthode de séquençage Sanger dans les années 1970, récompensée par le prix Nobel de chimie en 1980.⁵³ Cette technique tire avantage de l'action inhibitrice des 2',3'-didésoxynucléotides triphosphate qui, lors d'une réaction de polymérisation par la polymérase d'ADN I, obligent la terminaison de l'élongation de la chaîne d'oligonucléotides.⁵³ Une variété de technologies de séquençage ont ensuite été développées.⁵⁴⁻⁵⁶ La technologie développée par Sanger a permis de séquencer un génome humain pour la première fois en 2001, ce qui a constitué une étape importante dans le domaine.^{57,58}

Le besoin de séquencer des génomes rapidement et à bas prix a mené à une deuxième révolution dans le domaine, soit des méthodes de séquençages de nouvelle génération (NGS pour « *Next Generation Sequencing* »). La technique Illumina a dominé cette seconde vague technologique, permettant le séquençage à haut débit de fragments d'oligonucléotides de quelques centaines de bases.⁵⁹ Ces fragments sont ensuite assemblés par des algorithmes pour reconstruire le génome d'intérêt.⁶⁰ Cette technique, qui ne nécessite ni clonage dans des vecteurs bactériens ni résolution de fragments d'ADN par électrophorèse, a permis de réduire les coûts de séquençage (Figure 1.3.).

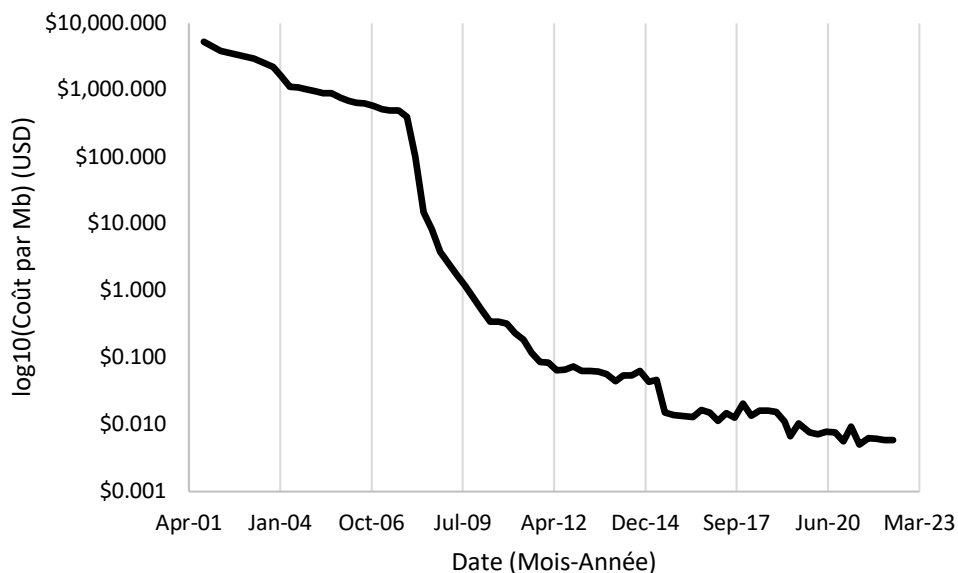


Figure 1.3. Le coût du séquençage a chuté depuis le début du séquençage à grande échelle. Les données proviennent du National Human Genome Research Institute (NHGRI).

Une des limitations inhérentes des méthodologies de NGS est la petite taille des fragments séquencés, ce qui peut résulter en des segments manquants (« *gaps* ») et des erreurs dans les séquences assemblées. La troisième et plus récente révolution dans les technologies de séquençage est définie comme le séquençage à longue lecture, ou de troisième génération (TGS pour « *Third Generation Sequencing* », ou « *Long-Read Sequencing* »).⁶¹ Cette technologie, marquée par le séquençage en temps réel et le séquençage par molécule unique (ou « *Single-Molecule Sequencing* »), permet de séquencer des fragments jusqu'à ~1 million de bases en longueur.⁶²⁻⁶⁴

Le développement des technologies de NGS, qui réduisent en particulier le coût et la quantité d'ADN nécessaire (Figure 1.3.), a donné naissance au domaine de la métagénomique, terme introduit pour la première fois en 1998.⁶⁵ Ce domaine, qui correspond à l'étude du matériel génomique provenant d'échantillons environnementaux, se distingue de la génomique du fait que l'ADN séquencé ne provient pas d'un organisme cultivé connu, mais d'une population complexe d'organismes. Ceci permet de séquencer des organismes qui ne peuvent pas être cultivés en laboratoire, par exemple en raison de leur étroite association avec d'autres espèces nécessaires à leur croissance.⁶⁶ À la différence des technologies NGS utilisées pour reconstituer un génome complet de plusieurs millions de bases, les lectures (« *reads* ») par les technologies de NGS en métagénomique sont assemblées en '*contigs*' (jargon de l'anglais, signifiant 'contigus') d'une longueur de quelques centaines de bases, jusqu'à quelques dizaines de milliers.^{67,68} En bref, à partir d'un échantillon environnemental, l'ADN est extrait, puis séquencé et assemblé en *contigs* qui

seront ensuite annotés.⁶⁹ Ces *contigs* appartiennent à une diversité d'organismes, tous présents dans l'échantillon de départ. Ainsi, puisque les échantillons métagénomiques comprennent une grande diversité de génomes en petite quantité, il n'est pas possible d'obtenir un recouvrement complet de chacun de ces génomes. La taille des *contigs* impacte la qualité des prédictions taxonomiques de ces séquences, et diminue le nombre de gènes qui peuvent être annotés sur un même contig. Ainsi, une des limitations majeures de la métagénomique est la quantité limitée d'information que peuvent offrir les séquences en résultant.

1.4.2 La prédiction de gènes

Alors que les génomes deviennent de plus en plus accessibles avec le développement des technologies de séquençage, il est nécessaire de développer des méthodologies avancées pouvant prédire les gènes qui s'y retrouvent. Rapidement, deux approches émergent. Les méthodes *ab initio* (aussi nommées intrinsèques), introduites dans les années 1980,⁷⁰⁻⁷² prédisent des gènes à partir de modèles statistiques qui utilisent l'information se trouvant à l'intérieur et à l'extérieur des régions codantes. Les méthodes par similarité (aussi nommées extrinsèques) utilisent plutôt des méthodes d'alignement local de séquences avec des bases de données de séquences de protéines ou de nucléotides de référence pour prédire les protéines putatives au sein de la séquence d'ADN d'intérêt.⁷³ Cette approche est dépendante de la qualité des bases de données, qui peuvent contenir des séquences de mauvaise qualité et peuvent ne pas contenir de séquences similaires à celles présentes dans le génome d'intérêt.⁷⁴ Ces deux approches ont rapidement été combinées pour permettre l'annotation de génomes complets.⁷⁵

Le défi de la prédiction des cadres de lecture n'est pas simple. Les génomes d'eucaryotes, riches en introns et régions non codantes, diffèrent de manière importante de ceux des procaryotes, qui comportent des régions codantes denses se chevauchant.⁷⁶ De plus, les génomes des procaryotes diffèrent entre groupements taxonomiques. Leurs tailles, leurs compositions en gènes, leur utilisation de codons et leur contenu en GC font en sorte que des outils de prédiction offrent une performance différente selon l'espèce étudiée.⁷⁷

Ainsi, il existe plusieurs séries d'opérations informatiques, dites 'pipelines informatiques' pour faire l'annotation de génomes, avec des niveaux de performance variables. La combinaison des deux approches est toujours d'actualité : le *National Center for Biotechnology Information* (NCBI), dont la base de données RefSeq contient plus de 300 millions de séquences protéiques en 2024, inclut toujours des algorithmes de prédiction *ab initio* et d'annotation par similarité.^{78,79}

En particulier, l'annotation de séquences métagénomiques est un défi, puisqu'elles comportent un mélange de séquences provenant d'eucaryotes, de procaryotes et de virus, qui doivent être annotées via différents pipelines.⁸⁰ Pourtant, l'annotation des gènes au sein des génomes et des séquences métagénomiques est une

étape critique de l'analyse de ces séquences. Le contexte génomique, qui se définit par l'organisation des gènes au sein d'un génome, est une source d'information importante pour l'étude de la fonction de gènes et de leur évolution. Par exemple, si l'identité des gènes voisins d'un gène d'intérêt est conservée, on peut inférer la présence d'un opéron, et ainsi la fonction du gène d'intérêt.^{81,82} Également, la présence de certains marqueurs géniques à proximité d'un gène d'intérêt, tels que des transposons et des intégrons cliniques, peut être un indicateur de transfert horizontal de gènes – un type d'évolution, où ces structures géniques permettent la mobilisation de gènes entre organismes à proximité.

1.4.3 Le développement de bases de données

Dès le début des années 1980, la nécessité d'établir des bases de données où sont entreposées les données de séquençage s'impose.⁸³ Leur rôle principal est de rendre les données disponibles, à la fois rapidement et le plus largement possible, et fait donc appel à un immense besoin de collaboration entre tous les acteurs du domaine. Parmi les défis qui sont apparus à travers les années, on retrouve le partage et la mobilité des données, la standardisation des entrées, le partage d'information entre les bases de données de différents pays – telles que GenBank aux États-Unis, EMBL-Bank en Europe et DDBJ au Japon – et l'inclusion de séquences nucléotidiques de plus en plus longues.⁸⁴

À présent, il existe plusieurs bases de données contenant des séquences nucléotidiques et protéiques qui répondent à différents besoins. Ces bases de données peuvent être composées de collections de bases de données, et avoir différentes versions. Les bases de données d'UniProt (pour « *Universal Protein Resource* »), UniProtKB, UniRef et UniParc, contiennent la majorité des séquences protéiques publiquement disponibles, et représentent une ressource centrale importante d'annotations variées relatives aux protéines (Figure 1.4).⁸⁵ Ce genre de base de données extrêmement vaste permet de faire des méta-analyses établissant des tendances au sein de l'espace de séquences.⁸⁶⁻⁸⁸

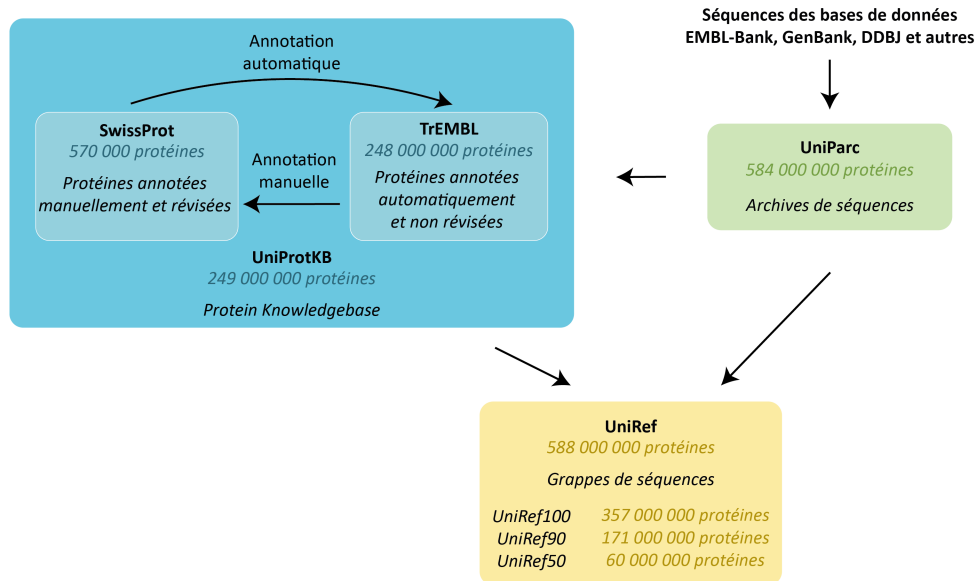


Figure 1.4. Les bases de données protéiques d’UniProt comportent plusieurs collections et versions.

Les protéines prédites provenant de multiples bases de données génomiques sont incluses dans UniParc. Ensuite, les séquences non redondantes sont intégrées dans TrEMBL et annotées automatiquement. Les séquences révisées pour lesquelles de l’information fonctionnelle et expérimentale est disponible sont intégrées dans SwissProt. UniRef contient des grappes de séquences venant de UniProtKB et UniParc, avec différents niveaux de redondance. Figure inspirée de <https://www.uniprot.org/help/about>.

Alors que la taille des bases de données augmente (Figure 1.5), la curation de ces bases de données est nécessaire. La redondance, soit la présence répétée d’informations similaires ou identiques, est un enjeu majeur puisqu’elle augmente de manière importante la taille des bases de données, et mène à la surreprésentation de certaines séquences. Ainsi, les gestionnaires des bases de données doivent prendre des décisions pour minimiser à la fois la redondance et la perte d’information. Depuis 2015, UniProt identifie les séquences presque identiques qui proviennent d’une même espèce et place les doublons dans UniParc.⁸⁹ Il existe également plusieurs versions du jeu de données UniRef, qui regroupent des ensembles de séquences à différents niveaux d’identité de séquence, soit à 50, 90 et 100% (Figure 1.4). Ainsi, ces différentes versions permettent de regrouper des séquences représentatives de tout l’espace de séquences exploré, avec différents degrés de recouvrement.

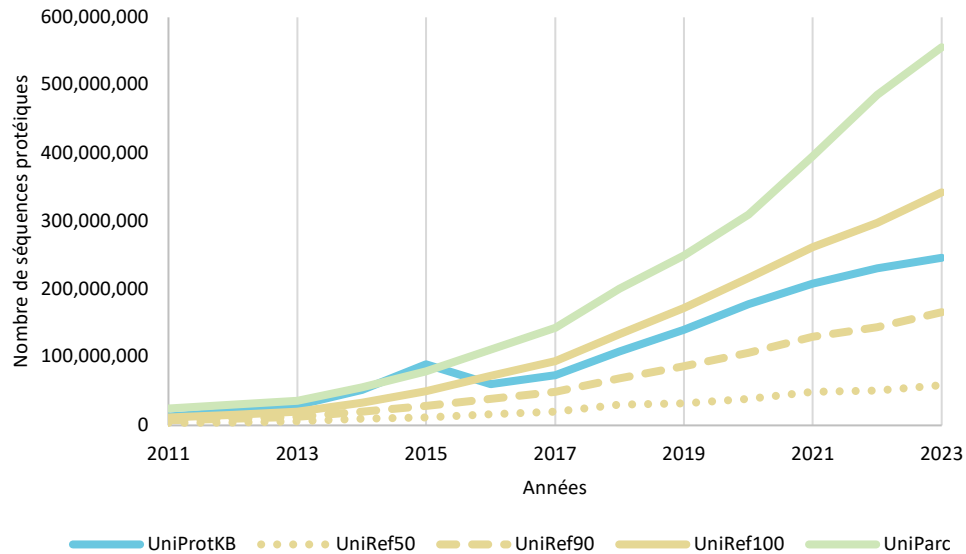


Figure 1.5. Croissance des bases de données d’UniProt depuis 2011.

Puisque les bases de données les plus étendues et largement utilisées sont nourries par des contributeurs divers avec des intérêts de recherche particuliers, elles ne sont pas une représentation fidèle de l’espace de séquence des protéines; elles comportent un biais pour certains types d’organismes et d’environnements qui sont majoritairement séquencés – en particulier, ce qui concerne l’être humain et les organismes qui l’affectent, ainsi que certains organismes modèles. Si la recherche qu’on vise faire s’intéresse à des éléments d’un type d’environnement particulier, comme son protéome ou son abondance quantitative de protéines, il est possible d’accéder à des bases de données se concentrant sur certains environnements particuliers⁹⁰ ou de générer son propre jeu de données⁹¹.

1.4.4 Outils bio-informatiques pour prédire la fonction de protéines

L’annotation de la fonction d’une protéine est centrale à plusieurs études. Puisque la majorité (>99.9%) des séquences protéiques d’UniProtKB sont putatives, le recours à des outils de prédiction de fonction est inévitable. En revanche, cette prédiction est particulièrement difficile.⁹² D’abord, la fonction d’une protéine peut concerner plusieurs niveaux, soit le rôle biochimique lui-même, mais également comment chaque protéine influence d’autres protéines, des voies métaboliques, la cellule, le tissu et l’organisme dans lequel elle se retrouve. Notre connaissance des fonctions des protéines nous vient de leur caractérisation expérimentale. Cette caractérisation implique rarement la description de tout son répertoire fonctionnel ; elle se fait généralement sous des conditions particulières, où l’étendue des conditions possibles, telles que la température et le pH, n’est pas testée. De plus, puisque les protéines ont souvent des activités promiscuitaires²² et/ou sont multifonctionnelles⁹, il est complexe de décrire et de connaître l’étendue des

fonctions réelles d'une protéine. Finalement, les annotations manuelles incomplètes et erronées peuvent diminuer la qualité des prédictions. Présentement, la base de données *Gene Ontology* (GO) est la source d'information la plus complète concernant la fonction de gènes.⁹³ Fondée en 1998, elle décrit trois aspects fonctionnels des protéines : la fonction moléculaire (ex. '*ATP hydrolysis binding*'), la composante cellulaire (ex. '*nucleoplasm*') et le processus biologique (ex. '*DNA replication*').

Plusieurs approches sont utilisées pour la prédiction fonctionnelle. La prédiction peut se faire par homologie, alors qu'on pose l'hypothèse que la similarité de séquence corrèle avec la similarité de fonction.⁹² La dernière décennie a permis à l'intelligence artificielle de s'immiscer dans les approches de prédiction de fonction pour les rendre plus précises. DeepFRI (pour « *Deep Functional Residue Identification* ») est l'une des méthodes les plus utilisées à ce jour.⁹⁴ Elle se base sur les réseaux convolutifs graphiques (ou « *graph convolutional networks* ») qui constituent un type d'architecture d'apprentissage profond, qui est lui-même un type d'intelligence artificielle. À partir de la séquence et de la structure d'une protéine, DeepFRI génère les probabilités pour chaque fonction prédite et identifie les résidus importants pour ces fonctions. Pour l'instant, la précision et la fiabilité de DeepFRI ont été évaluées en comparant les prédictions fonctionnelles aux fonctions expérimentalement confirmées pour des chaînes PDB, ou encore en comparant les prédictions à celles basées sur la séquence uniquement.^{94,95}

Il est nécessaire de souligner que les méthodes de prédiction se basent sur les informations connues pour générer des prédictions. Ainsi, elles ne peuvent pas servir à investiguer des fonctions à ce jour inconnues chez les protéines, ou encore les fonctions des protéines qui sont structurellement et évolutivement non reliées aux protéines annotées. Pourtant, les techniques de prédiction actuelles hallucinent (inventent) souvent des fonctions lorsqu'on leur présente des séquences sans homologues.⁸⁷ La communauté scientifique déploie actuellement des efforts pour diminuer les prédictions erronées afin de favoriser l'annotation '*Uncharacterized protein*' par rapport à l'annotation de fonctions infondées, basées sur des annotations ambiguës.⁸⁷

1.4.5 Outils bio-informatiques pour identifier des relations évolutives entre protéines

Alors que la quantité de séquences accessible explose, il devient alors possible de comparer ces séquences pour étudier les processus d'évolution moléculaire. En 1970, Saul B. Needleman et Christian D. Wunsch publient un algorithme permettant de quantifier la similarité de deux séquences protéiques en utilisant un score de similarité.⁹⁶ À partir d'une matrice considérant toutes les combinaisons d'alignements possibles et attribuant un score aux substitutions, délétions et insertions, cet algorithme identifie l'alignement optimal global. En 1981, Temple F. Smith et Michael S. Waterman introduisent un algorithme quantifiant l'homologie locale entre deux séquences.⁹⁷ Cet algorithme vise à identifier la section de la plus similaire entre elles en effectuant une recherche exhaustive de tous les alignements locaux possibles. En comparaison

à l'alignement fait par l'algorithme de Needleman-Wunsch, les régions qui ne s'alignent pas de manière optimale n'influencent pas le score l'alignement pour la région où une homologie de séquence est détectée. Cela rend l'algorithme Smith-Waterman plus robuste, et son utilisation demeure une composante essentielle des pipelines informatiques explorant l'espace de séquences.

Ces algorithmes impliquent une utilisation intensive de ressources informatiques, et ne peuvent pas être généralisés à la comparaison d'une grande quantité de séquences. Ainsi, il est clair qu'une autre approche est nécessaire pour pouvoir détecter les séquences homologues à une séquence d'intérêt à travers de larges bases de données. À cet effet, en 1990, l'approche BLAST (pour « *B*asic *L*ocal *A*lignment *S*earch *T*ool »), a été publiée (Figure 1.6A).⁹⁸ En bref, à partir de la séquence soumise, BLAST la fractionne en de courtes sous-séquences (qualifiées de mot, ou « *seed* »), identifie des mots similaires et cherche chacun de ces mots à travers une base de données. Lorsque des séquences sont identifiées comme partageant ce mot, l'alignement est étendu à partir de ce mot pour identifier la région homologue entre les deux séquences. Un score de similarité pour l'alignement généré est établi par l'utilisation d'une matrice de notation des substitutions. Cette approche heuristique permet de faire des alignements de séquences très rapidement, au sacrifice de la sensibilité caractéristique de l'algorithme de Smith-Waterman. Ceci en fait l'un des outils les plus utilisés pour identifier des homologues.

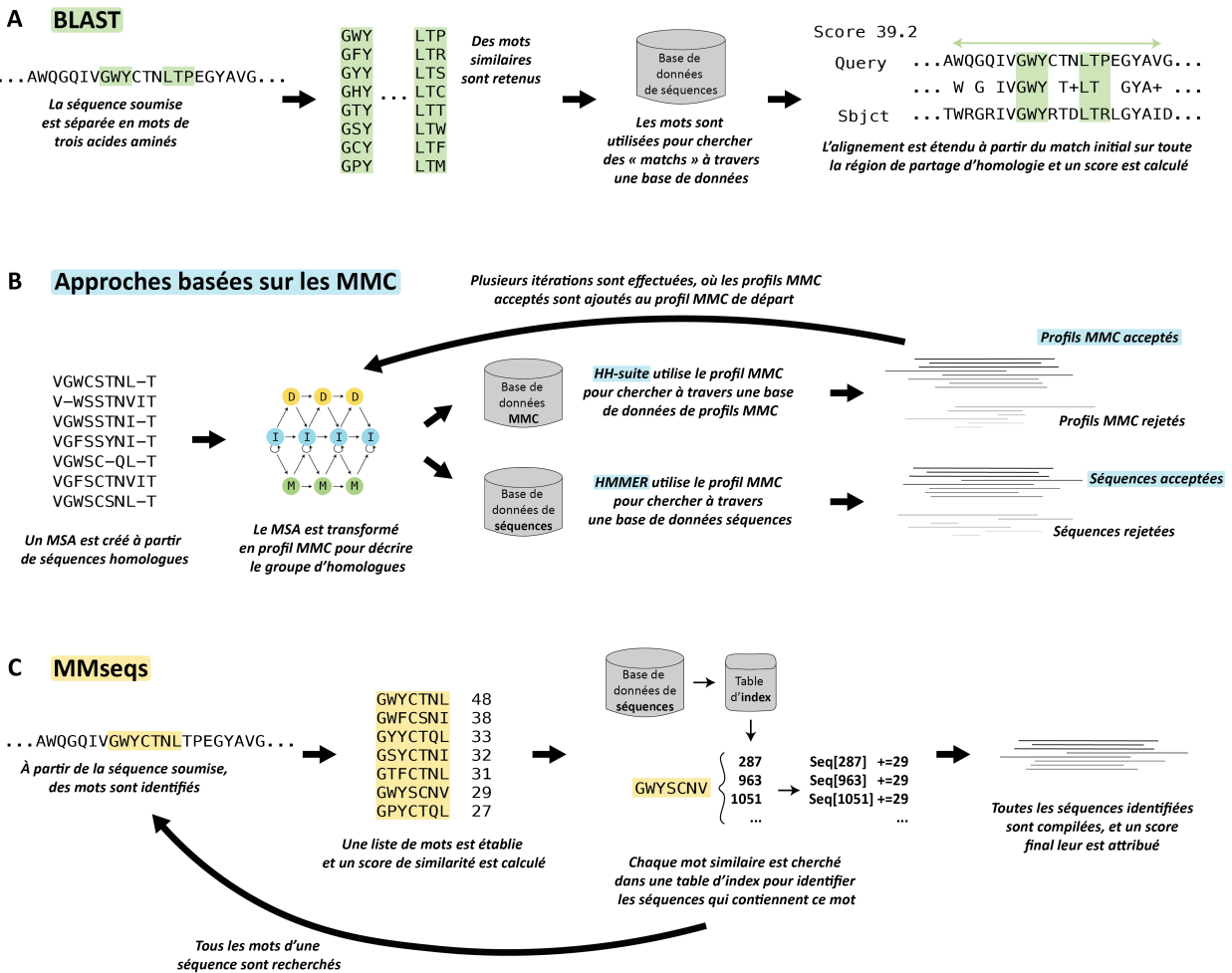


Figure 1.6. Plusieurs approches ont été développées pour identifier des homologues de séquences à partir d'une ou plusieurs séquences d'intérêt.

A. À partir d'une séquence soumise, BLAST sépare la séquence en mots, composés ici de trois acides aminés. Tous les mots présents dans la séquence sont identifiés ; ici, on en représente que deux. Des mots similaires à ceux de la séquence soumise, qui diffèrent par un acide aminé, sont identifiés et cherchés dans une base de données sélectionnée. Lorsqu'il y a au moins deux matches identifiés dans une séquence, l'alignement avec la séquence initiale est étendu et un score y est attribué. **B.** Les approches de HH-suite et HMMER sont représentées. Un profil de modèles de Markov cachés (MMC) est généré à partir de l'alignement de séquences multiples soumises. HH-suite identifie des profils MMC provenant d'une base de données de référence et partageant de l'homologie avec le MMC généré. Ces profils sont incorporés au profil MMC initial, permettant de mettre plus de poids aux régions de la séquence évolutivement conservées. Un score est attribué à tous les profils MMC identifiés pour départager les profils partageant une forte homologie avec le profil de départ de ceux avec une faible homologie. À partir du profil MMC généré, HMMER identifie des séquences partageant une homologie avec le profil à partir d'une base de données de séquences, et donne un score aux séquences identifiées. **C.** MMseqs identifie tous les mots de six ou sept acides aminés à partir de la séquence soumise. Une liste de mots similaires est générée, et un score de similarité au mot initial leur est donné. Les mots sont cherchés au sein d'un index préalablement généré pour identifier les séquences contenant ces mots. Ce processus est répété pour tous les mots. Les séquences identifiées qui contiennent au moins deux mots sont alignées à la séquence initiale, et un score est donné pour ces alignements.

Les algorithmes présentés ci-haut sont tous basés sur l'utilisation de matrices comparant des séquences, où l'importance de chaque position au sein de la séquence est considérée comme étant équivalente au reste de la séquence. Pourtant, il est admis que certaines régions d'une séquence sont essentielles à la fonction d'une macromolécule, et donc conservées, alors que d'autres régions sont sujettes aux substitutions, insertions et délétions. L'utilisation d'un profil, soit une représentation statistique d'une famille de séquences apparentées dérivée d'un alignement multiple de séquences (MSA pour « *Multiple Sequence Alignment* »), permet de décrire les séquences en prenant en compte cette réalité. Les modèles de Markov cachés (MMC, connu sous le terme HMM pour « *Hidden Markov Model* ») permettent de définir une distribution de probabilité sur un nombre infini de séquences possibles (Figure 1.7).⁹⁹ En bref, une chaîne de Markov décrit une séquence d'évènements. Dans ce modèle stochastique, la probabilité du prochain évènement est conditionnée par l'état de l'évènement actuel.⁹⁹ Les MMC sont des modèles statistiques de Markov où l'on assume que le système est une chaîne de Markov comportant des états observables Y qui sont eux-mêmes influencés par les états cachés X . Les protéines, qui sont ultimement une chaîne d'acides aminés, peuvent être décrites par ces modèles ; dans ce cas-ci, l'état caché X correspond à la fonction biologique, alors qu'on ne peut seulement observer l'état de séquence Y .¹⁰⁰ L'idée est d'apprendre sur l'état caché X – la fonction – à partir de l'état observable Y – la séquence. Ainsi, les profils MMC décrivant des familles de protéines impliquent la création d'un modèle probabiliste basé sur l'alignement multiple des séquences. Pfam¹¹, maintenant disponible via InterPro¹⁰¹, est une base de données de familles de protéines répertoriées, chacune décrite par un profil MMC, qui peut être utilisée pour identifier les homologues d'une même famille protéique.

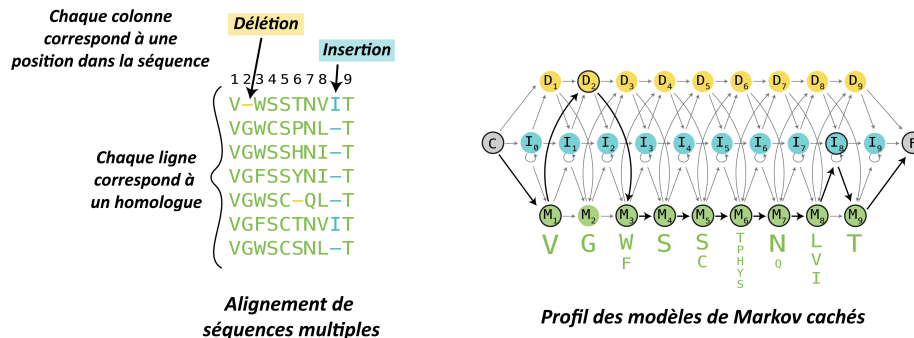


Figure 1.7. Plusieurs représentations peuvent être utilisées pour décrire une famille de protéines.

À gauche est représenté un alignement de séquences multiples de sept séquences homologues. Certaines positions, où le même acide aminé est observé au sein de toute la colonne, sont très conservées. Pour d'autres positions, on peut observer la délétion ou l'insertion d'un résidu. À droite est illustré un profil des modèles de Markov cachés (MMC). L'état M correspond à une position dans la séquence. La fréquence d'un résidu au sein d'une colonne du MSA est exprimée en probabilité au sein de cet état (représentée ici par la taille des lettres en vert). Les états I et D décrivent la probabilité d'observer une insertion et une délétion, respectivement, à une position donnée. Les états observables de la première séquence du MSA sont illustrés sur le profil MMC ; il commence à l'état de commencement (C), puis suit les flèches noires épaisses, jusqu'à l'état final (F).

Ces profils MMC sont non seulement une façon efficace pour décrire des familles de protéines, mais sont également une manière sophistiquée d'explorer l'espace de séquences. Par contre, tout comme l'algorithme de Smith-Waterman, les approches par MMC utilisent des méthodes de programmation dynamique qui ont un fort besoin en ressources de calcul. Ainsi, l'algorithme HMMER, développé par Sean Eddy, a implanté un algorithme d'accélération heuristique à la recherche par MMC, le rendant plus efficace que BLAST (Figure 1.6B).^{102,103} L'algorithme HH-suite est également utilisé pour faire des recherches d'homologie de séquences en se basant sur les MMC (Figure 1.6B).¹⁰⁴ La différence principale entre HMMER et HH-suite est que le premier crée des profils MMC et cherche les occurrences de ces profils dans une base de données de séquences cibles, alors que le deuxième effectue des recherches itératives de comparaison où les profils MMC d'intérêts sont comparés à une base de données de profils MMC.

Récemment, MMseqs (pour « *Many-against-Many sequence searching* ») s'est imposé comme algorithme ultrarapide de recherche, puisqu'il est optimisé pour travailler avec de très grandes bases de données (Figure 1.6C).¹⁰⁵ D'abord, pour créer une base de données de référence que MMseqs pourra utiliser efficacement, cet outil regroupe les séquences protéiques d'une base de données partageant une homologie de séquence ; chaque groupe est ensuite indexé pour faciliter la manipulation des données. Lorsqu'une séquence est soumise à MMseqs, les groupes de séquences étant les plus similaires sont récupérés grâce leur index, et les mots provenant de la séquence soumise sont recherchés pour identifier des séquences homologues. Pour toutes les séquences homologues potentielles, MMseqs effectue des alignements de séquences avec l'algorithme de Smith-Waterman et calcule les scores de similarité. Les séquences ayant un score supérieur au seuil établi sont ensuite rapportées.

1.4.6 Prédiction de structures et complexes protéiques

Alors que les séquences protéiques sont à la base de toute étude sur leur évolution, la détermination de leur structure tridimensionnelle est un aspect fondamental pour la pleine compréhension de la fonction des protéines. La *Protein Data Bank* (PDB) est la doyenne des bases de données biologiques ; elle archive depuis 1971 les structures de macromolécules biologiques.¹⁰⁶ Cependant, en raison de la complexité inhérente à la résolution expérimentale des structures de macromolécules, son expansion n'est pas aussi remarquable que celle des bases de données de séquences. Après plus de 50 années d'existence, la PDB comporte présentement un peu plus de 200 000 structures, dont plusieurs sont codées par des séquences identiques ou quasi-identiques, représentant 0.08% des séquences se trouvant dans UniProtKB. Ainsi, l'espace de séquence couvert par les structures expérimentalement résolues est infime, largement influencé par des limitations expérimentales.

Dès la fin des années 1990s, il est apparu évident que la prédiction de structures allait être un acteur clé dans notre compréhension de l'espace de structures de protéines exploré naturellement. En 1994, la

compétition CASP (pour « *Critical Assessment of Structure Prediction* ») a vu le jour¹⁰⁷ ; il s'agit d'une compétition bisannuelle où la communauté scientifique prédit des structures de macromolécules selon divers niveaux de difficulté. Les prédictions sont comparées à des structures expérimentalement résolues – qui ne sont pas accessibles préalablement à la compétition – ce qui permet de faire le point sur les meilleures approches de prédiction du moment.

Ainsi, depuis près de 30 ans, plusieurs méthodes ont été employées pour résoudre le problème de la prédiction de structures à partir d'une séquence protéique. Traditionnellement, deux approches sont utilisées.¹⁰⁸ Les méthodes basées sur les modèles (« *template-based* » ou encore « *homology modeling* ») se servent de structures résolues pour prédire la structure de protéines dont la séquence est similaire. En bref, ces méthodes identifient une (ou plusieurs) séquence(s) homologue(s) dont la structure est connue, y alignent la séquence d'intérêt et modélisent les différences correspondant aux substitutions, insertions et délétions par rapport au modèle de départ. D'un autre côté, les méthodes sans modèles (*ab initio* ou « *template-free* ») se basent sur un MSA pour prédire les structures locales et des contacts au sein de la chaîne d'acides aminés, suivie par une stratégie d'échantillonnage conformationnel pour générer des modèles. Ces modèles sont ensuite affinés, classés et comparés aux autres pour définir les meilleurs.

En 2020, au CASP14, l'algorithme AlphaFold2 (AF2) a non seulement révolutionné le domaine de la prédiction de structure, mais également tout le domaine de la biologie.¹⁰⁹⁻¹¹¹ AF2 utilise une approche hybride, intégrant à la fois des éléments des méthodes basées sur les modèles et celles *ab initio*. En bref, à partir de la séquence pour laquelle on veut prédire une structure, AF2 crée un MSA en cherchant des homologues à travers de vastes bases de données, telles que les bases de données métagénomiques Mgnify¹¹² et BFD¹¹³, avec HMMER et HH-suite. AF2 utilise ensuite le module Evoformer, une architecture d'apprentissage profond appelée réseau neuronal, pour prédire la distance dans l'espace entre les acides aminés de la séquence soumise à partir du MSA. En particulier, les signaux de coévolution présents dans le MSA, aussi appelés corrélation ou covariations, sont critiques pour la prédiction ; lorsque les résidus d'une colonne d'un MSA varient conjointement avec les résidus d'une autre colonne, il est possible que ces résidus interagissent dans l'espace (Figure 1.8).¹¹⁴ Ce module intègre également les informations relatives aux structures dont la séquence est similaire à celle dont on veut prédire la structure pour affiner ses prédictions. AF2 prédit ensuite les angles du squelette peptidique des sous-sections de la séquence, assemble la structure, affine ses modèles itérativement et donne ensuite un score à sa prédiction pour informer de son niveau de confiance.

L'approche décrite ici pour prédire les interactions au sein d'une chaîne protéique permet la prédiction d'interactions entre différentes chaînes, qui est assurée par l'algorithme AlphaFold-multimer.¹¹⁵ La toute récente *AlphaFold Protein Structure Database* (AlphaFold DB), regroupant actuellement plus de

210 000 000 de structures issues d'entrées UniProt standards, ou non virales, marque une avancée extraordinaire en matière de données dans le domaine de la biologie structurale.¹¹⁶

Les avancées des dernières décennies, telles que la démocratisation de technologies de séquençage, le développement de larges bases de données métagénomiques, la création d'outils de recherche de séquences basées sur les MMC et le développement des approches en intelligence artificielle, ont été instrumentales au développement d'AF2.

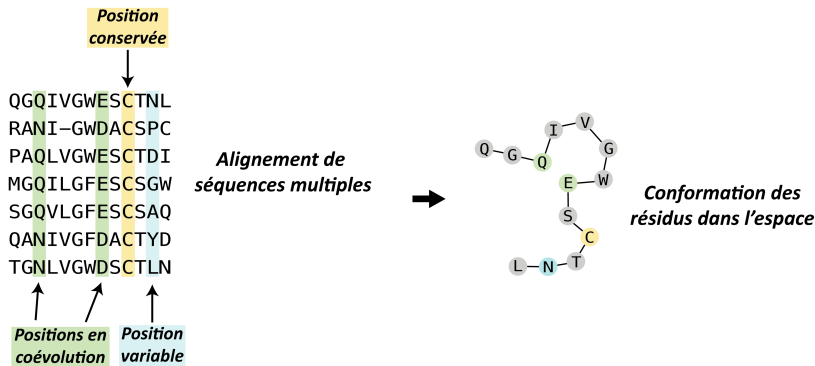


Figure 1.8. Les alignements de séquences multiples (MSA) informent sur la structure des protéines dans l'espace.

À gauche : un MSA est présenté. La colonne indiquée en jaune correspond à une position conservée, possiblement pour son importance structurelle ou fonctionnelle, alors que la colonne indiquée en bleu correspond à une position variable, dont l'identité de la chaîne latérale n'est pas importante pour la fonction. Les deux colonnes surlignées en vert varient conjointement ; soit l'identité de cette paire de résidus est de Gln et Glu, soit elle est de Asn et Asp. À droite : l'arrangement tridimensionnel de la première séquence du MSA est représenté. Les résidus dont les positions varient conjointement dans le MSA sont à proximité dans l'espace.

Il est important de noter que pour l'instant, aucune approche n'est basée uniquement sur les principes physiques guidant le repliement d'une chaîne d'acides aminés dans l'espace dans son état le plus stable. Ceci met en évidence les limites de notre compréhension des lois de la nature qui font qu'une structure protéique primaire se replie dans l'état qui représente l'énergie libre la plus faible, comme l'a postulé Christian B. Anfinsen il y a 50 ans.⁴⁹

1.5 Utilisation de la technologie pour l'étude de l'évolution catalytique

La combinaison des technologies présentées ci-haut révolutionne le domaine de la biochimie évolutive. Il est maintenant possible de reconstituer de manière précise les chemins évolutifs empruntés¹¹⁷ ou encore d'identifier des relations évolutives entre des enzymes et repliements protéiques apparemment distincts¹¹⁸. Ainsi, ces avancées permettent de former un pont entre les domaines de la biochimie – l'étude des propriétés chimiques et physiques des molécules biologiques – et de la biologie évolutive – l'étude des mécanismes évolutifs qui ont mené à la diversité des organismes vivants.⁸⁶

Les efforts modernes en biochimie évolutive démontrent que pour avoir une compréhension approfondie des facteurs à l'œuvre lors de l'évolution d'enzymes, il est essentiel d'avoir une vue holistique du système d'intérêt en utilisant des approches en génomique, en biologie structurale, en bio-informatique et en biologie moléculaire. En effet, les approches récentes permettant d'établir l'historique évolutif de familles de protéines intègrent les informations relatives à la séquence, à la structure, au contexte génomique, à la prédiction de fonction et aux résultats expérimentaux fonctionnels.

Par exemple, en reconstruisant la phylogénie des réductases de ribonucléotide, Burnim et ses collègues ont identifié une sous-famille éloignée sur le plan évolutif, qui représente l'ancêtre de deux des trois grandes sous-familles modernes.¹¹⁹ Ils ont résolu expérimentalement et computationnellement la structure de cette sous-famille, la qualifiant comme étant la forme la plus minimale des réductases de ribonucléotide connues à ce jour. Au sein de tous les opérons où cette sous-famille est présente, que ce soit chez des bactéries ou des phages, un gène de type ferritine est systématiquement situé immédiatement en aval du gène de la réductase de ribonucléotide, démontrant ainsi sa conservation sur le plan du contexte génomique. En combinant l'ensemble de ces données, l'équipe a pu élaborer un modèle évolutif pour comprendre l'évolution de la famille des réductases de ribonucléotide.

Dans un autre exemple, en adoptant une démarche à grande échelle, Durairaj et son équipe ont répertorié toutes les protéines se trouvant dans la base de données AlphaFold DB et ont recueilli des données fonctionnelles disponibles pour chacune d'entre elles.⁸⁷ L'équipe a exploré en détail certaines familles d'enzymes isolées évolutivement et ne comportant aucune annotation fonctionnelle. Par exemple, pour une famille comptant 159 séquences représentatives annotées comme DUF6516 (*Domain of Unknown Function 6516*), l'algorithme DeepFRI a prédit un site actif pour l'hydrolase de liaisons ester, dont la fonction précise était inconnue. L'analyse du contexte génomique de ces séquences a révélé la présence d'un gène conservé en amont, qui s'est avéré être homologue à une antitoxine et a été prédit, par DeepFRI, pour avoir une affinité de liaison à l'ADN, une caractéristique commune aux antitoxines. En outre, les prédictions d'AlphaFold-multimer ont suggéré que ces deux partenaires interagissent sous forme de dimères de dimères. L'équipe a donc émis l'hypothèse que ces deux séquences formeraient un système toxine-antitoxine, où DUF6516 serait la toxine. Cette hypothèse a été expérimentalement validée : la famille jusqu'alors de fonction inconnue a été identifiée comme étant un nouvel effecteur toxique ciblant la traduction. Cela illustre le succès d'une approche interdisciplinaire pour caractériser une famille de protéines nouvellement identifiée. Cette approche holistique revêt une importance significative dans le contexte des familles de protéines qui sont fortement isolées sur le plan de la séquence et qui présentent peu ou pas de similitude avec d'autres protéines déjà connues.

1.5.1 Le modèle de la dihydrofolate réduction de type B

Dès mon arrivée au doctorat en 2019, je me suis intéressée à une famille d'enzymes largement étudiée sur les plans biophysique et enzymatique : la dihydrofolate réductase de type B (DfrB). Ces enzymes ont été découvertes dans les années 1970 puisqu'elles confèrent une résistance à l'antibiotique synthétique triméthopime (TMP), introduit en clinique quelques années auparavant.¹²⁰⁻¹²³ Alors que la dihydrofolate réductase ubiquitaire bactérienne (FolA) est inhibée sélectivement par le TMP, les DfrB – qui sont structurellement et évolutivement non reliées aux FolA malgré leur activité catalytique commune – ne le sont pas, leur permettant de produire le cofacteur essentiel tétrahydrofolate, même en présence de cet antibiotique.¹²⁴ Par contre, le mécanisme d'émergence récente de cette famille d'enzymes résistantes au triméthopime reste encore inexpliqué.

En effet, depuis leur découverte, les recherches entreprises sur les DfrB se sont attardées aux caractéristiques qui les distinguent des enzymes typiques. Ma directrice de thèse et moi-même avons publié une revue de littérature décrivant en détail les connaissances actuelles concernant cette famille d'enzymes dans la revue *Chemical Communications*, qui constitue le Chapitre 2 de cette thèse. En bref, quatre protomères sont nécessaires pour former un homotétramère qui constitue la forme active de l'enzyme (Figure 1.9).

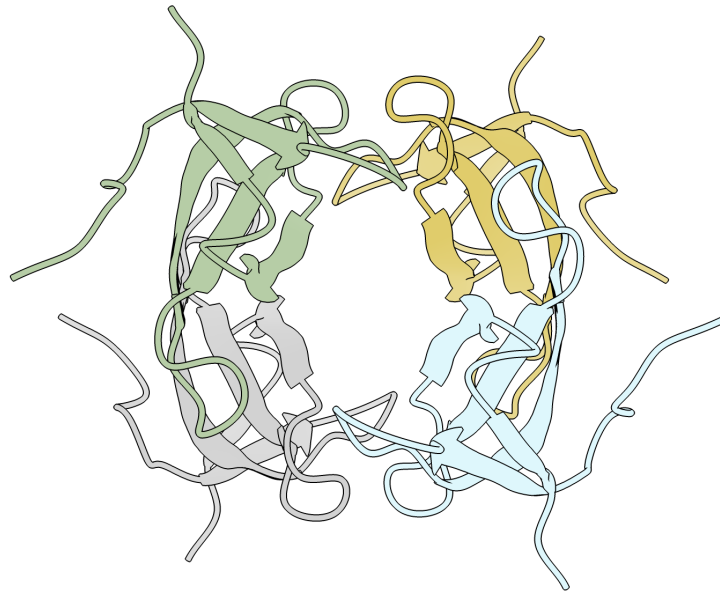


Figure 1.9. Structure de la DfrB1, une enzyme homotétramérique (PDB 2rk1).

Chaque monomère, composé du repliement SH3, est indiqué par une couleur différente. Le pore central constitue le site actif unique de l'enzyme. Les chaînes latérales des quatre résidus du site actif, sur chacun des protomères, sont représentées en bâtonnets. Figure réalisée avec ChimeraX (<https://www.rbvi.ucsf.edu/chimerax>).

Le tunnel central est l'unique site actif de l'enzyme ; les DfrB sont les seules enzymes connues pour lesquelles quatre protomères contribuent à la formation d'un seul site actif. Ces enzymes catalysent la réduction du dihydrofolate en utilisant le NADPH comme cofacteur réducteur. Le substrat et le cofacteur entrent chacun par une des deux entrées identiques du tunnel. Le transfert d'hydrure entre le groupement nicotinamide du NADPH et du groupement ptérine du dihydrofolate se fait au centre du pore central, alors que les résidus du site actif orientent les groupements chimiques nécessaires à cette transformation (tel que présenté en plus de détail au Chapitre 2). Contrairement à ce qui est décrit pour la vaste majorité des enzymes, le laboratoire Pelletier a démontré qu'aucun des résidus du site actif des DfrB n'est essentiel à son mécanisme de catalyse par proximité.¹²⁵ Un autre aspect curieux des DfrB est leur faible efficacité catalytique, qui équivaut à moins d'un point de pourcentage de celle des Fola. Ainsi, elles ne semblent pas optimisées pour la réduction du dihydrofolate, à la différence des Fola.

Les DfrB ont fait l'objet d'études biophysiques et enzymatiques poussées pour comprendre leur fonctionnement, puisqu'elles se distinguent significativement des enzymes typiques. Curieusement, les DfrB ne partagent pas d'homologie de séquence avec d'autres protéines caractérisées. Ainsi, une question fondamentale les concernant reste entière : comment ces enzymes ont-elles émergé dans le contexte récent de la résistance aux antibiotiques ?

Une hypothèse plausible est que les DfrB ont évolué une activité dihydrofolate réductase suite à l'introduction du triméthopime dans les années 1960s. La pression sélective exercée sur les bactéries sujettes au triméthopime aurait favorisé, auprès d'un ancêtre des DfrB, la sélection de mutations qui lui aurait ultimement permis de lier le substrat dihydrofolate et le cofacteur NADPH et catalyser le transfert d'hydrure entre ces derniers, produisant le cofacteur essentiel à la prolifération cellulaire, le tétrahydrofolate. Une telle évolution rapide d'une nouvelle activité catalytique suite à l'introduction d'un composé synthétique a déjà été démontrée.¹²⁶ Cette hypothèse forme le fondement des chapitres expérimentaux de cette thèse, soient les chapitres 3 à 5.

1.5.2 Efforts pour établir l'origine des DfrB

Depuis leur découverte, les DfrB ont été rapportées à quelques reprises au sein d'éléments génétiques mobiles (MGE pour « *Mobile Genetic Element* ») de bactéries pathogènes aux humains et aux animaux.¹²⁷⁻¹²⁹ Parmi les MGE, on retrouve les plasmides, qui peuvent s'échanger entre bactéries, les intégrons, qui facilitent l'échange de cassettes de gènes et les transposons, qui permettent à des segments d'ADN d'être transposés entre chromosomes, bactériophages ou plasmides.¹³⁰ Jusqu'en 2019, les DfrB étaient exclusivement identifiées au sein de ces structures. On ignorait donc l'origine des DfrB préalablement à leur intégration dans le résistome mobile, correspondant aux gènes codant pour les mécanismes de résistance aux antibiotiques.

Dans le but d'obtenir une meilleure vue d'ensemble sur le contexte génomique dans lequel se trouvent les DfrB, j'ai d'abord cherché ces gènes dans les bases de données génomiques publiquement disponibles. Ces travaux, présentés au Chapitre 3, identifient les DfrB comme se trouvant non seulement dans les plasmides mais également dans les chromosomes de bactéries pathogènes, et rapportent pour la première fois un gène des DfrB dans un contexte génomique différent de celui relié à la résistance aux antibiotiques, soit dans un organisme environnemental.

Ensuite, pour obtenir des indices concernant l'origine des DfrB, j'ai entrepris l'identification et la caractérisation de leurs homologues à l'aide de la base de données UniProtKB grâce à leur appartenance à la même famille MMC. Ces travaux, présentés au Chapitre 4, ont démontré que les homologues des DfrB présentant la même capacité à procurer une résistance au triméthoprim se retrouvent dans des organismes environnementaux divers, n'ayant donc subi aucune pression sélective par l'exposition au triméthoprim. Également, ces travaux ont démontré que le domaine des DfrB peut être fusionné à d'autres domaines protéiques tout en conservant sa capacité à créer un tétramère pouvant catalyser la réduction du dihydrofolate.

Fort de notre succès à identifier des homologues fonctionnellement identiques aux DfrB, j'ai entrepris l'exploration poussée de l'espace de séquence du domaine DfrB en parcourant une diversité des bases de données génomiques et métagénomiques avec les technologies de recherche de séquences par MMC. Ces recherches m'ont permis de délimiter les frontières fonctionnelles de ce domaine, en mettant en lumière les caractéristiques évolutives qui ont influencé son évolution. J'ai accompli cela, entre autres, en caractérisant expérimentalement des homologues par le biais de criblage pour la fonction de la résistance au triméthoprim, et en utilisant l'outil de prédiction de structure AlphaFold-multimer ; ces efforts sont présentés au Chapitre 5.

Par conséquent, mes recherches ont réfuté l'hypothèse la plus probable concernant l'évolution des DfrB, selon laquelle une évolution de la fonction enzymatique aurait suivi l'introduction de l'antibiotique triméthoprim. Dans le Chapitre 5, je présente un modèle évolutif pour le domaine des DfrB et je soutiens dans le chapitre de discussion que l'évolution découlant de l'introduction du triméthoprim en clinique relève davantage de l'ordre génomique.

En somme, au sein des Chapitres 3 à 5, j'ai tiré avantage des avancées technologiques récentes pour informer sur l'émergence du domaine DfrB. Les avancées en séquençage, les diverses technologies pour explorer l'espace de séquence et celles pour prédire de manière fiable la structure de complexes protéiques ont été essentielles à ce succès.

1.6 Le prochain défi technologique

Mon approche pour étudier l'émergence des DfrB ne tient pas compte de la dynamique moléculaire du système. Cette omission s'explique par les résultats des études en résonance magnétique nucléaire (RMN) et en dynamique moléculaire (MD, pour « Molecular Dynamics »), qui ont révélé la rigidité du squelette peptidique de la DfrB1.^{131,132} Ces études ont aussi montré qu'aucun changement structural n'était observé lors de la liaison des substrats de l'enzyme. Ainsi, la dynamique protéique n'est pas un élément qui définit la catalyse pour ce système.

Ceci dit, la dynamique protéique est un élément central à l'évolution de nombreux systèmes enzymatiques.¹³³ En raison de la complexité que nécessite la génération et l'analyse des données de dynamique, la dynamique protéique ne fait pas encore partie des pipelines de caractérisation de systèmes enzymatiques. Pourtant, les efforts de la communauté scientifique vont dans ce sens. Ainsi, depuis quelques années, de plus en plus de travaux de recherche caractérisent la dynamique protéique de variants enzymatiques et d'homologues, ce qui permet de mieux comprendre le rôle que joue la dynamique protéique dans l'évolution et la modification de l'activité d'enzymes. En Annexe 1, je présente un compte-rendu des efforts actuels pour intégrer les connaissances sur la dynamique des enzymes au processus d'ingénierie des enzymes, ce qui consoliderait notre compréhension de cette propriété inhérente aux protéines et à leur évolution.

1.7 Références

- (1) Longo, L. M.; Petrović, D.; Kamerlin, S. C. L.; Tawfik, D. S. Short and Simple Sequences Favored the Emergence of N-Helix Phospho-Ligand Binding Sites in the First Enzymes. *Proc. Natl. Acad. Sci.* **2020**, *117* (10), 5310–5318. <https://doi.org/10.1073/pnas.1911742117>.
- (2) Edbeib, M. F.; Wahab, R. A.; Huyop, F. Halophiles: Biology, Adaptation, and Their Role in Decontamination of Hypersaline Environments. *World J. Microbiol. Biotechnol.* **2016**, *32* (8), 135. <https://doi.org/10.1007/s11274-016-2081-9>.
- (3) Feller, G. Protein Stability and Enzyme Activity at Extreme Biological Temperatures. *J. Phys. Condens. Matter* **2010**, *22* (32), 323101. <https://doi.org/10.1088/0953-8984/22/32/323101>.
- (4) Hedstrom, L. Enzyme Specificity and Selectivity. In *eLS*; John Wiley & Sons, Ltd, Ed.; Wiley, 2010. <https://doi.org/10.1002/9780470015902.a0000716.pub2>.
- (5) Tawfik, O. K. and D. S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annu. Rev. Biochem.* **2010**, *79* (1), 471–505. <https://doi.org/10.1146/annurev-biochem-030409-143718>.
- (6) Tokuriki, N.; Tawfik, D. S. Protein Dynamism and Evolvability. *Science* **2009**, *324* (5924), 203–207. <https://doi.org/10.1126/science.1169375>.
- (7) Kirschner, K.; Bisswanger, H. Multifunctional Proteins. *Annu. Rev. Biochem.* **1976**, *45*, 143–166. <https://doi.org/10.1146/annurev.bi.45.070176.001043>.
- (8) Copley, S. D. Moonlighting Is Mainstream: Paradigm Adjustment Required. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **2012**, *34* (7), 578–588. <https://doi.org/10.1002/bies.201100191>.

- (9) Jeffery, C. J. Moonlighting Proteins. *Trends Biochem. Sci.* **1999**, *24* (1), 8–11. [https://doi.org/10.1016/S0968-0004\(98\)01335-8](https://doi.org/10.1016/S0968-0004(98)01335-8).
- (10) Kelley, L. A.; Sternberg, M. J. Partial Protein Domains: Evolutionary Insights and Bioinformatics Challenges. *Genome Biol.* **2015**, *16* (1), 100. <https://doi.org/10.1186/s13059-015-0663-8>.
- (11) Finn, R. D.; Bateman, A.; Clements, J.; Coghill, P.; Eberhardt, R. Y.; Eddy, S. R.; Heeger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E. L. L.; Tate, J.; Punta, M. Pfam: The Protein Families Database. *Nucleic Acids Res.* **2014**, *42* (D1), D222–D230. <https://doi.org/10.1093/nar/gkt1223>.
- (12) Kolodny, R.; Nepomnyachiy, S.; Tawfik, D. S.; Ben-Tal, N. Bridging Themes: Short Protein Segments Found in Different Architectures. *Mol. Biol. Evol.* **2021**, *38* (6), 2191–2208. <https://doi.org/10.1093/molbev/msab017>.
- (13) Hadley, C.; Jones, D. T. A Systematic Comparison of Protein Structure Classifications: SCOP, CATH and FSSP. *Structure* **1999**, *7* (9), 1099–1112. [https://doi.org/10.1016/S0969-2126\(99\)80177-4](https://doi.org/10.1016/S0969-2126(99)80177-4).
- (14) Cheng, H.; Schaeffer, R. D.; Liao, Y.; Kinch, L. N.; Pei, J.; Shi, S.; Kim, B.-H.; Grishin, N. V. ECoD: An Evolutionary Classification of Protein Domains. *PLoS Comput. Biol.* **2014**, *10* (12), e1003926. <https://doi.org/10.1371/journal.pcbi.1003926>.
- (15) Cordes, M. H.; Davidson, A. R.; Sauer, R. T. Sequence Space, Folding and Protein Design. *Curr. Opin. Struct. Biol.* **1996**, *6* (1), 3–10. [https://doi.org/10.1016/S0959-440X\(96\)80088-1](https://doi.org/10.1016/S0959-440X(96)80088-1).
- (16) Clifton, B. E.; Kozome, D.; Laurino, P. Efficient Exploration of Sequence Space by Sequence-Guided Protein Engineering and Design. *Biochemistry* **2023**, *62* (2), 210–220. <https://doi.org/10.1021/acs.biochem.1c00757>.
- (17) Atkinson, H. J.; Morris, J. H.; Ferrin, T. E.; Babbitt, P. C. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLoS ONE* **2009**, *4* (2), e4345. <https://doi.org/10.1371/journal.pone.0004345>.
- (18) Copp, J. N.; Akiva, E.; Babbitt, P. C.; Tokuriki, N. Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry* **2018**, *57* (31), 4651–4662. <https://doi.org/10.1021/acs.biochem.8b00473>.
- (19) Jensen, R. A. Enzyme Recruitment in Evolution of New Function. *Annu. Rev. Microbiol.* **1976**, *30* (1), 409–425. <https://doi.org/10.1146/annurev.mi.30.100176.002205>.
- (20) Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Gould, S. M.; Roodveldt, C.; Tawfik, D. S. The “evolvability” of Promiscuous Protein Functions. *Nat. Genet.* **2005**, *37* (1), 73–76. <https://doi.org/10.1038/ng1482>.
- (21) Copley, S. D. An Evolutionary Biochemist’s Perspective on Promiscuity. *Trends Biochem. Sci.* **2015**, *40* (2), 72–78. <https://doi.org/10.1016/j.tibs.2014.12.004>.
- (22) Khersonsky, O.; Roodveldt, C.; Tawfik, D. Enzyme Promiscuity: Evolutionary and Mechanistic Aspects. *Curr. Opin. Chem. Biol.* **2006**, *10* (5), 498–508. <https://doi.org/10.1016/j.cbpa.2006.08.011>.
- (23) Zou, T.; Risso, V. A.; Gavira, J. A.; Sanchez-Ruiz, J. M.; Ozkan, S. B. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Mol. Biol. Evol.* **2015**, *32* (1), 132–143. <https://doi.org/10.1093/molbev/msu281>.
- (24) Pandya, C.; Farelli, J. D.; Dunaway-Mariano, D.; Allen, K. N. Enzyme Promiscuity: Engine of Evolutionary Innovation. *J. Biol. Chem.* **2014**, *289* (44), 30229–30236. <https://doi.org/10.1074/jbc.R114.572990>.

- (25) Palmer, D. R. J.; Garrett, J. B.; Sharma, V.; Meganathan, R.; Babbitt, P. C.; Gerlt, J. A. Unexpected Divergence of Enzyme Function and Sequence: “*N*-Acylamino Acid Racemase” Is *o*-Succinylbenzoate Synthase. *Biochemistry* **1999**, *38* (14), 4252–4258. <https://doi.org/10.1021/bi990140p>.
- (26) Noda-Garcia, L.; Tawfik, D. S. Enzyme Evolution in Natural Products Biosynthesis: Target- or Diversity-Oriented? *Curr. Opin. Chem. Biol.* **2020**, *59*, 147–154. <https://doi.org/10.1016/j.cbpa.2020.05.011>.
- (27) Todd, A. E.; Orengo, C. A.; Thornton, J. M. Sequence and Structural Differences between Enzyme and Nonenzyme Homologs. *Structure* **2002**, *10* (10), 1435–1451. [https://doi.org/10.1016/S0969-2126\(02\)00861-4](https://doi.org/10.1016/S0969-2126(02)00861-4).
- (28) Tam, R.; Saier, M. H. A Bacterial Periplasmic Receptor Homologue with Catalytic Activity: Cyclohexadienyl Dehydratase of *Pseudomonas Aeruginosa* Is Homologous to Receptors Specific for Polar Amino Acids. *Res. Microbiol.* **1993**, *144* (3), 165–169. [https://doi.org/10.1016/0923-2508\(93\)90041-Y](https://doi.org/10.1016/0923-2508(93)90041-Y).
- (29) Ngaki, M. N.; Louie, G. V.; Philippe, R. N.; Manning, G.; Pojer, F.; Bowman, M. E.; Li, L.; Larsen, E.; Wurtele, E. S.; Noel, J. P. Evolution of the Chalcone-Isomerase Fold from Fatty-Acid Binding to Stereospecific Catalysis. *Nature* **2012**, *485* (7399), 530–533. <https://doi.org/10.1038/nature11009>.
- (30) Ortmayer, M.; Lafite, P.; Menon, B. R. K.; Tralau, T.; Fisher, K.; Denkhaus, L.; Scrutton, N. S.; Rigby, S. E. J.; Munro, A. W.; Hay, S.; Leys, D. An Oxidative N-Demethylase Reveals PAS Transition from Ubiquitous Sensor to Enzyme. *Nature* **2016**, *539* (7630), 593–597. <https://doi.org/10.1038/nature20159>.
- (31) Kaltenbach, M.; Burke, J. R.; Dindo, M.; Pabis, A.; Munsberg, F. S.; Rabin, A.; Kamerlin, S. C. L.; Noel, J. P.; Tawfik, D. S. Evolution of Chalcone Isomerase from a Noncatalytic Ancestor. *Nat. Chem. Biol.* **2018**, *14* (6), 548–555. <https://doi.org/10.1038/s41589-018-0042-3>.
- (32) Clifton, B. E.; Kaczmarek, J. A.; Carr, P. D.; Gerth, M. L.; Tokuriki, N.; Jackson, C. J. Evolution of Cyclohexadienyl Dehydratase from an Ancestral Solute-Binding Protein. *Nat. Chem. Biol.* **2018**, *14* (6), 542–547. <https://doi.org/10.1038/s41589-018-0043-2>.
- (33) Wang, X.; Minasov, G.; Shoichet, B. K. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-Offs. *J. Mol. Biol.* **2002**, *320* (1), 85–95. [https://doi.org/10.1016/S0022-2836\(02\)00400-X](https://doi.org/10.1016/S0022-2836(02)00400-X).
- (34) Wu, G.; Fiser, A.; Ter Kuile, B.; Šali, A.; Müller, M. Convergent Evolution of *Trichomonas Vaginalis* Lactate Dehydrogenase from Malate Dehydrogenase. *Proc. Natl. Acad. Sci.* **1999**, *96* (11), 6285–6290. <https://doi.org/10.1073/pnas.96.11.6285>.
- (35) Beintema, J. J.; Schüller, C.; Irie, M.; Carsana, A. Molecular Evolution of the Ribonuclease Superfamily. *Prog. Biophys. Mol. Biol.* **1988**, *51* (3), 165–192. [https://doi.org/10.1016/0079-6107\(88\)90001-6](https://doi.org/10.1016/0079-6107(88)90001-6).
- (36) Medeiros, A. A. Evolution and Dissemination of β -Lactamases Accelerated by Generations of β -Lactam Antibiotics. *Clin. Infect. Dis.* **1997**, *24* (Supplement_1), S19–S45. https://doi.org/10.1093/clinids/24.Supplement_1.S19.
- (37) Perona, J. J.; Craik, C. S. Evolutionary Divergence of Substrate Specificity within the Chymotrypsin-like Serine Protease Fold. *J. Biol. Chem.* **1997**, *272* (48), 29987–29990. <https://doi.org/10.1074/jbc.272.48.29987>.
- (38) Bystroff, C.; Oatley, S. J.; Kraut, J. Crystal Structures of *Escherichia Coli* Dihydrofolate Reductase: The NADP⁺ Holoenzyme and the Folate .Cndtdot. NADP⁺ Ternary Complex. Substrate Binding and

- a Model for the Transition State. *Biochemistry* **1990**, *29* (13), 3263–3277. <https://doi.org/10.1021/bi00465a018>.
- (39) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. Crystal Structures of *Escherichia Coli* and *Lactobacillus Casei* Dihydrofolate Reductase Refined at 1.7 Å Resolution. I. General Features and Binding of Methotrexate. *J. Biol. Chem.* **1982**, *257* (22), 13650–13662. [https://doi.org/10.1016/S0021-9258\(18\)33497-5](https://doi.org/10.1016/S0021-9258(18)33497-5).
- (40) Davies, J. F.; Delcamp, T. J.; Prendergast, N. J.; Ashford, V. A.; Freisheim, J. H.; Kraut, J. Crystal Structures of Recombinant Human Dihydrofolate Reductase Complexed with Folate and 5-Deazafolate. *Biochemistry* **1990**, *29* (40), 9467–9479. <https://doi.org/10.1021/bi00492a021>.
- (41) McTigue, M. A.; Davies, J. F.; Kaufman, B. T.; Kraut, J. Crystal Structures of Chicken Liver Dihydrofolate Reductase: Binary thioNADP⁺ and Ternary thioNADP⁺.Cntdot.Biopterin Complexes. *Biochemistry* **1993**, *32* (27), 6855–6862. <https://doi.org/10.1021/bi00078a008>.
- (42) Sawaya, M. R.; Kraut, J. Loop and Subdomain Movements in the Mechanism of *Escherichia Coli* Dihydrofolate Reductase: Crystallographic Evidence. *Biochemistry* **1997**, *36* (3), 586–603. <https://doi.org/10.1021/bi962337c>.
- (43) Bhabha, G.; Ekiert, D. C.; Jennewein, M.; Zmasek, C. M.; Tuttle, L. M.; Kroon, G.; Dyson, H. J.; Godzik, A.; Wilson, I. A.; Wright, P. E. Divergent Evolution of Protein Conformational Dynamics in Dihydrofolate Reductase. *Nat. Struct. Mol. Biol.* **2013**, *20* (11), 1243–1249. <https://doi.org/10.1038/nsmb.2676>.
- (44) Bönisch, H.; Backmann, J.; Kath, T.; Naumann, D.; Schäfer, G. Adenylate Kinase from *Sulfolobus Acidocaldarius*: Expression in *Escherichia Coli* and Characterization by Fourier Transform Infrared Spectroscopy. *Arch. Biochem. Biophys.* **1996**, *333* (1), 75–84. <https://doi.org/10.1006/abbi.1996.0366>.
- (45) Lacher, K.; Schafer, G. Archaeobacterial Adenylate Kinase from the Thermoacidophile *Sulfolobus Acidocaldarius*: Purification, Characterization, and Partial Sequence. *Arch. Biochem. Biophys.* **1993**, *302* (2), 391–397. <https://doi.org/10.1006/abbi.1993.1229>.
- (46) Haney, P. J.; Stees, M.; Konisky, J. Analysis of Thermal Stabilizing Interactions in Mesophilic and Thermophilic Adenylate Kinases from the Genus *Methanococcus*. *J. Biol. Chem.* **1999**, *274* (40), 28453–28458. <https://doi.org/10.1074/jbc.274.40.28453>.
- (47) Haney, P.; Konisky, J.; Koretke, K. K.; Luthey-Schulten, Z.; Wolynes, P. G. Structural Basis for Thermostability and Identification of Potential Active Site Residues for Adenylate Kinases from the Archaeal genus *Methanococcus*. *Proteins Struct. Funct. Genet.* **1997**, *28* (1), 117–130. [https://doi.org/10.1002/\(SICI\)1097-0134\(199705\)28:1<117::AID-PROT12>3.0.CO;2-M](https://doi.org/10.1002/(SICI)1097-0134(199705)28:1<117::AID-PROT12>3.0.CO;2-M).
- (48) Sela, M.; White, F. H.; Anfinsen, C. B. Reductive Cleavage of Disulfide Bridges in Ribonuclease. *Science* **1957**, *125* (3250), 691–692. <https://doi.org/10.1126/science.125.3250.691>.
- (49) Anfinsen, C. B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181* (4096), 223–230. <https://doi.org/10.1126/science.181.4096.223>.
- (50) Chothia, C.; Lesk, A. M. The Relation between the Divergence of Sequence and Structure in Proteins. *EMBO J.* **1986**, *5* (4), 823–826. <https://doi.org/10.1002/j.1460-2075.1986.tb04288.x>.
- (51) Ohno, S. *Evolution by Gene Duplication*; Springer Berlin Heidelberg: Berlin, Heidelberg, 1970. <https://doi.org/10.1007/978-3-642-86659-3>.
- (52) Neurath, H.; Walsh, K. A.; Winter, W. P. Evolution of Structure and Function of Proteases: Amino Acid Sequences of Proteolytic Enzymes Reflect Phylogenetic Relationships. *Science* **1967**, *158* (3809), 1638–1644. <https://doi.org/10.1126/science.158.3809.1638>.

- (53) Sanger, F.; Nicklen, S.; Coulson, A. R. DNA Sequencing with Chain-Terminating Inhibitors. *Proc. Natl. Acad. Sci.* **1977**, *74* (12), 5463–5467. <https://doi.org/10.1073/pnas.74.12.5463>.
- (54) Nyrén, P.; Lundin, A. Enzymatic Method for Continuous Monitoring of Inorganic Pyrophosphate Synthesis. *Anal. Biochem.* **1985**, *151* (2), 504–509. [https://doi.org/10.1016/0003-2697\(85\)90211-8](https://doi.org/10.1016/0003-2697(85)90211-8).
- (55) Maxam, A. M.; Gilbert, W. A New Method for Sequencing DNA. *Proc. Natl. Acad. Sci.* **1977**, *74* (2), 560–564. <https://doi.org/10.1073/pnas.74.2.560>.
- (56) Ohara, R.; Ohara, O. A New Solid-Phase Chemical DNA Sequencing Method Which Uses Streptavidin-Coated Magnetic Beads. *DNA Res.* **1995**, *2* (3), 123–128. <https://doi.org/10.1093/dnares/2.3.123>.
- (57) International Human Genome Sequencing Consortium; Whitehead Institute for Biomedical Research, Center for Genome Research;; Lander, E. S.; Linton, L. M.; Birren, B.; Nusbaum, C.; Zody, M. C.; Baldwin, J.; Devon, K.; Dewar, K.; et al. Initial Sequencing and Analysis of the Human Genome. *Nature* **2001**, *409* (6822), 860–921. <https://doi.org/10.1038/35057062>.
- (58) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; et al. The Sequence of the Human Genome. *Science* **2001**, *291* (5507), 1304–1351. <https://doi.org/10.1126/science.1058040>.
- (59) Quail, M. A.; Kozarewa, I.; Smith, F.; Scally, A.; Stephens, P. J.; Durbin, R.; Swerdlow, H.; Turner, D. J. A Large Genome Center’s Improvements to the Illumina Sequencing System. *Nat. Methods* **2008**, *5* (12), 1005–1010. <https://doi.org/10.1038/nmeth.1270>.
- (60) Miller, J. R.; Koren, S.; Sutton, G. Assembly Algorithms for Next-Generation Sequencing Data. *Genomics* **2010**, *95* (6), 315–327. <https://doi.org/10.1016/j.ygeno.2010.03.001>.
- (61) Van Dijk, E. L.; Jaszczyszyn, Y.; Naquin, D.; Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **2018**, *34* (9), 666–681. <https://doi.org/10.1016/j.tig.2018.05.008>.
- (62) Rhoads, A.; Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **2015**, *13* (5), 278–289. <https://doi.org/10.1016/j.gpb.2015.08.002>.
- (63) Manrao, E. A.; Derrington, I. M.; Laszlo, A. H.; Langford, K. W.; Hopper, M. K.; Gillgren, N.; Pavlenok, M.; Niederweis, M.; Gundlach, J. H. Reading DNA at Single-Nucleotide Resolution with a Mutant MspA Nanopore and Phi29 DNA Polymerase. *Nat. Biotechnol.* **2012**, *30* (4), 349–353. <https://doi.org/10.1038/nbt.2171>.
- (64) Jain, M.; Koren, S.; Miga, K. H.; Quick, J.; Rand, A. C.; Sasani, T. A.; Tyson, J. R.; Beggs, A. D.; Dilthey, A. T.; Fiddes, I. T.; et al. Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads. *Nat. Biotechnol.* **2018**, *36* (4), 338–345. <https://doi.org/10.1038/nbt.4060>.
- (65) Handelsman, J.; Rondon, M. R.; Brady, S. F.; Clardy, J.; Goodman, R. M. Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products. *Chem. Biol.* **1998**, *5* (10), R245–R249. [https://doi.org/10.1016/S1074-5521\(98\)90108-9](https://doi.org/10.1016/S1074-5521(98)90108-9).
- (66) Tringe, S. G.; Rubin, E. M. Metagenomics: DNA Sequencing of Environmental Samples. *Nat. Rev. Genet.* **2005**, *6* (11), 805–814. <https://doi.org/10.1038/nrg1709>.
- (67) Crampton-Platt, A.; Yu, D. W.; Zhou, X.; Vogler, A. P. Mitochondrial Metagenomics: Letting the Genes out of the Bottle. *GigaScience* **2016**, *5* (1), 15. <https://doi.org/10.1186/s13742-016-0120-y>.
- (68) MetaHIT Consortium; Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K. S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; et al. A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing. *Nature* **2010**, *464* (7285), 59–65. <https://doi.org/10.1038/nature08821>.

- (69) Thomas, T.; Gilbert, J.; Meyer, F. Metagenomics - a Guide from Sampling to Data Analysis. *Microb. Inform. Exp.* **2012**, *2* (1), 3. <https://doi.org/10.1186/2042-5783-2-3>.
- (70) Fickett, J. W. Recognition of Protein Coding Regions in DNA Sequences. *Nucleic Acids Res.* **1982**, *10* (17), 5303–5318. <https://doi.org/10.1093/nar/10.17.5303>.
- (71) Gribskov, M.; Devereux, J.; Burgess, R. R. The Codon Preference Plot: Graphic Analysis of Protein Coding Sequences and Prediction of Gene Expression. *Nucleic Acids Res.* **1984**, *12* (1Part2), 539–549. <https://doi.org/10.1093/nar/12.1Part2.539>.
- (72) Staden, R. Measurements of the Effects That Coding for a Protein Has on a DNA Sequence and Their Use for Finding Genes. *Nucleic Acids Res.* **1984**, *12* (1Part2), 551–567. <https://doi.org/10.1093/nar/12.1Part2.551>.
- (73) Robison, K.; Gilbert, W.; Church, G. M. Large Scale Bacterial Gene Discovery by Similarity Search. *Nat. Genet.* **1994**, *7* (2), 205–214. <https://doi.org/10.1038/ng0694-205>.
- (74) Al-Turaiki, I. M.; Mathkour, H.; Touir, A.; Hammami, S. Computational Approaches for Gene Prediction: A Comparative Survey. In *Informatics Engineering and Information Science*; Abd Manaf, A., Zeki, A., Zamani, M., Chuprat, S., El-Qawasmeh, E., Eds.; Communications in Computer and Information Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; Vol. 252, pp 14–25. https://doi.org/10.1007/978-3-642-25453-6_2.
- (75) Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J.-F.; Dougherty, B. A.; Merrick, J. M.; et al. Whole-Genome Random Sequencing and Assembly of *Haemophilus Influenzae* Rd. *Science* **1995**, *269* (5223), 496–512. <https://doi.org/10.1126/science.7542800>.
- (76) Goel, N.; Singh, S.; Aseri, T. C. A Review of Soft Computing Techniques for Gene Prediction. *ISRN Genomics* **2013**, *2013*, 1–8. <https://doi.org/10.1155/2013/191206>.
- (77) Dimonaco, N. J.; Aubrey, W.; Kenobi, K.; Clare, A.; Creevey, C. J. No One Tool to Rule Them All: Prokaryotic Gene Prediction Tool Annotations Are Highly Dependent on the Organism of Study. *Bioinformatics* **2022**, *38* (5), 1198–1207. <https://doi.org/10.1093/bioinformatics/btab827>.
- (78) O’Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufu, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Res.* **2016**, *44* (D1), D733–D745. <https://doi.org/10.1093/nar/gkv1189>.
- (79) Tatusova, T.; DiCuccio, M.; Badretdin, A.; Chetvernin, V.; Nawrocki, E. P.; Zaslavsky, L.; Lomsadze, A.; Pruitt, K. D.; Borodovsky, M.; Ostell, J. NCBI Prokaryotic Genome Annotation Pipeline. *Nucleic Acids Res.* **2016**, *44* (14), 6614–6624. <https://doi.org/10.1093/nar/gkw569>.
- (80) Scholz, M. B.; Lo, C.-C.; Chain, P. S. Next Generation Sequencing and Bioinformatic Bottlenecks: The Current State of Metagenomic Data Analysis. *Curr. Opin. Biotechnol.* **2012**, *23* (1), 9–15. <https://doi.org/10.1016/j.copbio.2011.11.013>.
- (81) Korbel, J. O.; Jensen, L. J.; Von Mering, C.; Bork, P. Analysis of Genomic Context: Prediction of Functional Associations from Conserved Bidirectionally Transcribed Gene Pairs. *Nat. Biotechnol.* **2004**, *22* (7), 911–917. <https://doi.org/10.1038/nbt988>.
- (82) Overbeek, R.; Fonstein, M.; D’Souza, M.; Pusch, G. D.; Maltsev, N. The Use of Gene Clusters to Infer Functional Coupling. *Proc. Natl. Acad. Sci.* **1999**, *96* (6), 2896–2901. <https://doi.org/10.1073/pnas.96.6.2896>.
- (83) Walgate, R. Europe Leads on Sequences. *Nature* **1982**, *296* (5858), 596–596. <https://doi.org/10.1038/296596a0>.

- (84) Stevens, H. Globalizing Genomics: The Origins of the International Nucleotide Sequence Database Collaboration. *J. Hist. Biol.* **2018**, *51* (4), 657–691. <https://doi.org/10.1007/s10739-017-9490-y>.
- (85) The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523–D531. <https://doi.org/10.1093/nar/gkac1052>.
- (86) Harms, M. J.; Thornton, J. W. Evolutionary Biochemistry: Revealing the Historical and Physical Causes of Protein Properties. *Nat. Rev. Genet.* **2013**, *14* (8), 559–571. <https://doi.org/10.1038/nrg3540>.
- (87) Durairaj, J.; Waterhouse, A. M.; Mets, T.; Brodiazhenko, T.; Abdullah, M.; Studer, G.; Tauriello, G.; Akdel, M.; Andreeva, A.; Bateman, A.; Tenson, T.; Hauryliuk, V.; Schwede, T.; Pereira, J. Uncovering New Families and Folds in the Natural Protein Universe. *Nature* **2023**, *622* (7983), 646–653. <https://doi.org/10.1038/s41586-023-06622-3>.
- (88) Koehler Leman, J.; Szczerbiak, P.; Renfrew, P. D.; Gligorijevic, V.; Berenberg, D.; Vatanen, T.; Taylor, B. C.; Chandler, C.; Janssen, S.; Pataki, A.; et al. Sequence-Structure-Function Relationships in the Microbial Protein Universe. *Nat. Commun.* **2023**, *14* (1), 2351. <https://doi.org/10.1038/s41467-023-37896-w>.
- (89) The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506–D515. <https://doi.org/10.1093/nar/gky1049>.
- (90) Sunagawa, S.; Coelho, L. P.; Chaffron, S.; Kultima, J. R.; Labadie, K.; Salazar, G.; Djahanschiri, B.; Zeller, G.; Mende, D. R.; Alberti, A.; et al. Structure and Function of the Global Ocean Microbiome. *Science* **2015**, *348* (6237), 1261359. <https://doi.org/10.1126/science.1261359>.
- (91) Kneis, D.; Lemay-St-Denis, C.; Cellier-Goetghebeur, S.; Elena, A. X.; Berendonk, T. U.; Pelletier, J. N.; Heß, S. Trimethoprim Resistance in Surface and Wastewater Is Mediated by Contrasting Variants of the *dfpB* Gene. *ISME J.* **2023**. <https://doi.org/10.1038/s41396-023-01460-7>.
- (92) Radivojac, P.; Clark, W. T.; Oron, T. R.; Schnoes, A. M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; et al. A Large-Scale Evaluation of Computational Protein Function Prediction. *Nat. Methods* **2013**, *10* (3), 221–227. <https://doi.org/10.1038/nmeth.2340>.
- (93) The Gene Ontology Consortium. The Gene Ontology Resource: 20 Years and Still GOing Strong. *Nucleic Acids Res.* **2019**, *47* (D1), D330–D338. <https://doi.org/10.1093/nar/gky1055>.
- (94) Gligorijević, V.; Renfrew, P. D.; Kosciolk, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; Xavier, R. J.; Knight, R.; Cho, K.; Bonneau, R. Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nat. Commun.* **2021**, *12* (1), 3168. <https://doi.org/10.1038/s41467-021-23303-9>.
- (95) Maranga, M.; Szczerbiak, P.; Bezshapkin, V.; Gligorijevic, V.; Chandler, C.; Bonneau, R.; Xavier, R. J.; Vatanen, T.; Kosciolk, T. Comprehensive Functional Annotation of Metagenomes and Microbial Genomes Using a Deep Learning-Based Method. *mSystems* **2023**, *8* (2), e01178-22. <https://doi.org/10.1128/msystems.01178-22>.
- (96) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48* (3), 443–453. [https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4).
- (97) Smith, T. F.; Waterman, M. S. Comparison of Biosequences. *Adv. Appl. Math.* **1981**, *2* (4), 482–489. [https://doi.org/10.1016/0196-8858\(81\)90046-4](https://doi.org/10.1016/0196-8858(81)90046-4).

- (98) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- (99) Eddy, S. R. Hidden Markov Models. *Curr. Opin. Struct. Biol.* **1996**, *6* (3), 361–365. [https://doi.org/10.1016/S0959-440X\(96\)80056-X](https://doi.org/10.1016/S0959-440X(96)80056-X).
- (100) Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14* (9), 755–763. <https://doi.org/10.1093/bioinformatics/14.9.755>.
- (101) Paysan-Lafosse, T.; Blum, M.; Chuguransky, S.; Grego, T.; Pinto, B. L.; Salazar, G. A.; Bileschi, M. L.; Bork, P.; Bridge, A.; Colwell, L.; et al. InterPro in 2022. *Nucleic Acids Res.* **2023**, *51* (D1), D418–D427. <https://doi.org/10.1093/nar/gkac993>.
- (102) Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7* (10), e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>.
- (103) Eddy, S. R. A New Generation of Homology Search Tools Based on Probabilistic Inference. In *Genome Informatics 2009*; PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.: Pacifico Yokohama, Japan, 2009; pp 205–211. https://doi.org/10.1142/9781848165632_0019.
- (104) Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. HH-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation. *BMC Bioinformatics* **2019**, *20* (1), 473. <https://doi.org/10.1186/s12859-019-3019-7>.
- (105) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35* (11), 1026–1028. <https://doi.org/10.1038/nbt.3988>.
- (106) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112* (3), 535–542. [https://doi.org/10.1016/S0022-2836\(77\)80200-3](https://doi.org/10.1016/S0022-2836(77)80200-3).
- (107) Madej, T.; Gibrat, J.-F.; Bryant, S. H. Threading a Database of Protein Cores. *Proteins Struct. Funct. Genet.* **1995**, *23* (3), 356–369. <https://doi.org/10.1002/prot.340230309>.
- (108) Kuhlman, B.; Bradley, P. Advances in Protein Structure Prediction and Design. *Nat. Rev. Mol. Cell Biol.* **2019**, *20* (11), 681–697. <https://doi.org/10.1038/s41580-019-0163-x>.
- (109) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- (110) Thornton, J. M.; Laskowski, R. A.; Borkakoti, N. AlphaFold Heralds a Data-Driven Revolution in Biology and Medicine. *Nat. Med.* **2021**, *27* (10), 1666–1669. <https://doi.org/10.1038/s41591-021-01533-0>.
- (111) Pereira, J.; Simpkin, A. J.; Hartmann, M. D.; Rigden, D. J.; Keegan, R. M.; Lupas, A. N. High-accuracy Protein Structure Prediction in CASP14. *Proteins Struct. Funct. Bioinforma.* **2021**, *89* (12), 1687–1699. <https://doi.org/10.1002/prot.26171>.
- (112) Richardson, L.; Allen, B.; Baldi, G.; Beracochea, M.; Bileschi, M. L.; Burdett, T.; Burgin, J.; Caballero-Pérez, J.; Cochrane, G.; Colwell, L. J.; et al. MGnify: The Microbiome Sequence Data Analysis Resource in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D753–D759. <https://doi.org/10.1093/nar/gkac1080>.
- (113) Steinegger, M.; Mirdita, M.; Söding, J. Protein-Level Assembly Increases Protein Sequence Recovery from Metagenomic Samples Manyfold. *Nat. Methods* **2019**, *16* (7), 603–606. <https://doi.org/10.1038/s41592-019-0437-4>.

- (114) Ekeberg, M.; Lövkvist, C.; Lan, Y.; Weigt, M.; Aurell, E. Improved Contact Prediction in Proteins: Using Pseudolikelihoods to Infer Potts Models. *Phys. Rev. E* **2013**, *87* (1), 012707. <https://doi.org/10.1103/PhysRevE.87.012707>.
- (115) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Židek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. *Protein Complex Prediction with AlphaFold-Multimer*; preprint; Bioinformatics, 2021. <https://doi.org/10.1101/2021.10.04.463034>.
- (116) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50* (D1), D439–D444. <https://doi.org/10.1093/nar/gkab1061>.
- (117) Avizemer, Z.; Martí-Gómez, C.; Hoch, S. Y.; McCandlish, D. M.; Fleishman, S. J. *Evolutionary Paths That Link Orthogonal Pairs of Binding Proteins*; preprint; In Review, 2023. <https://doi.org/10.21203/rs.3.rs-2836905/v1>.
- (118) Alvarez-Carreño, C.; Penev, P. I.; Petrov, A. S.; Williams, L. D. Fold Evolution before LUCA: Common Ancestry of SH3 Domains and OB Domains. *Mol. Biol. Evol.* **2021**, *38* (11), 5134–5143. <https://doi.org/10.1093/molbev/msab240>.
- (119) Burnim, A. A.; Spence, M. A.; Xu, D.; Jackson, C. J.; Ando, N. Comprehensive Phylogenetic Analysis of the Ribonucleotide Reductase Family Reveals an Ancestral Clade. *eLife* **2022**, *11*, e79790. <https://doi.org/10.7554/eLife.79790>.
- (120) Pattishall, K. H.; Acar, J.; Burchall, J. J.; Goldstein, F. W.; Harvey, R. J. Two Distinct Types of Trimethoprim-Resistant Dihydrofolate Reductase Specified by R-Plasmids of Different Compatibility Groups. *J. Biol. Chem.* **1977**, *252* (7), 2319–2323.
- (121) Fleming, M. P.; Datta, N.; Gruneberg, R. N. Trimethoprim Resistance Determined by R Factors. *BMJ* **1972**, *1* (5802), 726–728. <https://doi.org/10.1136/bmj.1.5802.726>.
- (122) Amyes, S. G. B.; Smith, J. T. R-Factor Trimethoprim Resistance Mechanism: An Insusceptible Target Site. *Biochem. Biophys. Res. Commun.* **1974**, *58* (2), 412–418. [https://doi.org/10.1016/0006-291X\(74\)90380-5](https://doi.org/10.1016/0006-291X(74)90380-5).
- (123) Noall, E. W. P.; Sowards, H. F. G.; Waterworth, P. M. Successful Treatment of a Case of Proteus Septicaemia. *BMJ* **1962**, *2* (5312), 1101–1102. <https://doi.org/10.1136/bmj.2.5312.1101>.
- (124) Howell, E. E. Searching Sequence Space: Two Different Approaches to Dihydrofolate Reductase Catalysis. *ChemBioChem* **2005**, *6* (4), 590–600. <https://doi.org/10.1002/cbic.200400237>.
- (125) Schmitzer, A. R.; Lépine, F.; Pelletier, J. N. Combinatorial Exploration of the Catalytic Site of a Drug-Resistant Dihydrofolate Reductase: Creating Alternative Functional Configurations. *Protein Eng. Des. Sel.* **2004**, *17* (11), 809–819. <https://doi.org/10.1093/protein/gzh090>.
- (126) Noor, S.; Taylor, M. C.; Russell, R. J.; Jermin, L. S.; Jackson, C. J.; Oakeshott, J. G.; Scott, C. Intramolecular Epistasis and the Evolution of a New Enzymatic Function. *PLoS ONE* **2012**, *7* (6), e39822. <https://doi.org/10.1371/journal.pone.0039822>.
- (127) Toulouse, J. L.; Edens, T. J.; Alejaldre, L.; Manges, A. R.; Pelletier, J. N. Integron-Associated DfrB4, a Previously Uncharacterized Member of the Trimethoprim-Resistant Dihydrofolate Reductase B Family, Is a Clinically Identified Emergent Source of Antibiotic Resistance. *Antimicrob. Agents Chemother.* **2017**, *61* (5), e02665-16, /aac/61/5/e02665-16.atom. <https://doi.org/10.1128/AAC.02665-16>.

- (128) Mokracka, J.; Koczura, R.; Kaznowski, A. Multiresistant Enterobacteriaceae with Class 1 and Class 2 Integrons in a Municipal Wastewater Treatment Plant. *Water Res.* **2012**, *46* (10), 3353–3363. <https://doi.org/10.1016/j.watres.2012.03.037>.
- (129) Xu, H.; Broersma, K.; Miao, V.; Davies, J. Class 1 and Class 2 Integrons in Multidrug-Resistant Gram-Negative Bacteria Isolated from the Salmon River, British Columbia. *Can. J. Microbiol.* **2011**, *57* (6), 460–467. <https://doi.org/10.1139/w11-029>.
- (130) Partridge, S. R.; Kwong, S. M.; Firth, N.; Jensen, S. O. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin. Microbiol. Rev.* **2018**, *31* (4), e00088-17, /cmr/31/4/e00088-17.atom. <https://doi.org/10.1128/CMR.00088-17>.
- (131) Alonso, H.; Gillies, M. B.; Cummins, P. L.; Bliznyuk, A. A.; Gready, J. E. Multiple Ligand-Binding Modes in Bacterial R67 Dihydrofolate Reductase. *J. Comput. Aided Mol. Des.* **2005**, *19* (3), 165–187. <https://doi.org/10.1007/s10822-005-3693-6>.
- (132) Pitcher, W. H.; DeRose, E. F.; Mueller, G. A.; Howell, E. E.; London, R. E. NMR Studies of the Interaction of a Type II Dihydrofolate Reductase with Pyridine Nucleotides Reveal Unexpected Phosphatase and Reductase Activity. *Biochemistry* **2003**, *42* (38), 11150–11160. <https://doi.org/10.1021/bi0349874>.
- (133) Tang, Q.-Y.; Kaneko, K. Dynamics-Evolution Correspondence in Protein Structures. *Phys. Rev. Lett.* **2021**, *127* (9), 098103. <https://doi.org/10.1103/PhysRevLett.127.098103>.

Chapitre 2. Revue de littérature sur la famille des DfrB

Préface

Dans ce chapitre, je présente les connaissances actuelles sur le fonctionnement de la famille des enzymes DfrB. J'y souligne les éléments qui en font des enzymes particulières, dont leur structure tétramérique symétrique, leur mécanisme catalytique et leur permissivité de substitutions au site actif. Également, je présente les évidences accumulées depuis leur découverte suggérant que ces enzymes sont primitives. En plus de résumer les connaissances sur les DfrB, ce chapitre propose un mécanisme d'évolution pour les DfrB ; d'un domaine de liaison, les DfrB auraient émergé une activité catalytique.

Ce chapitre est composé d'une revue de littérature publiée dans le journal *Chemical Communications* suite à une invitation de leur part. Des modifications mineures ont été apportées à la version incluse dans cette thèse. J'ai réalisé la revue de littérature et écrit la première version du manuscrit. Prof. Joelle N. Pelletier et moi-même avons ensuite édité le manuscrit et conçu les figures, que j'ai réalisées.

Article de revue 1. From a binding module to essential catalytic activity: how nature stumbled on a good thing

Claudèle Lemay-St-Denis^{1,2,3} et Joelle N. Pelletier^{1,2,3,4,*}

¹ Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, QC H3T 1J4, Canada

² PROTEO, The Québec Network for Research on Protein, Function, Engineering and Applications, Québec, QC G1V 0A6, Canada

³ CGCC, Center in Green Chemistry and Catalysis, Montréal, QC H3A 0B8, Canada

⁴ Chemistry Department, Université de Montréal, Montréal, QC H2V 0B3, Canada

*Correspondence: joelle.pelletier@umontreal.ca

Chemical Communications

DOI : [10.1039/d3cc04209j](https://doi.org/10.1039/d3cc04209j)

© The Royal Society of Chemistry 2023

2.1 Abstract

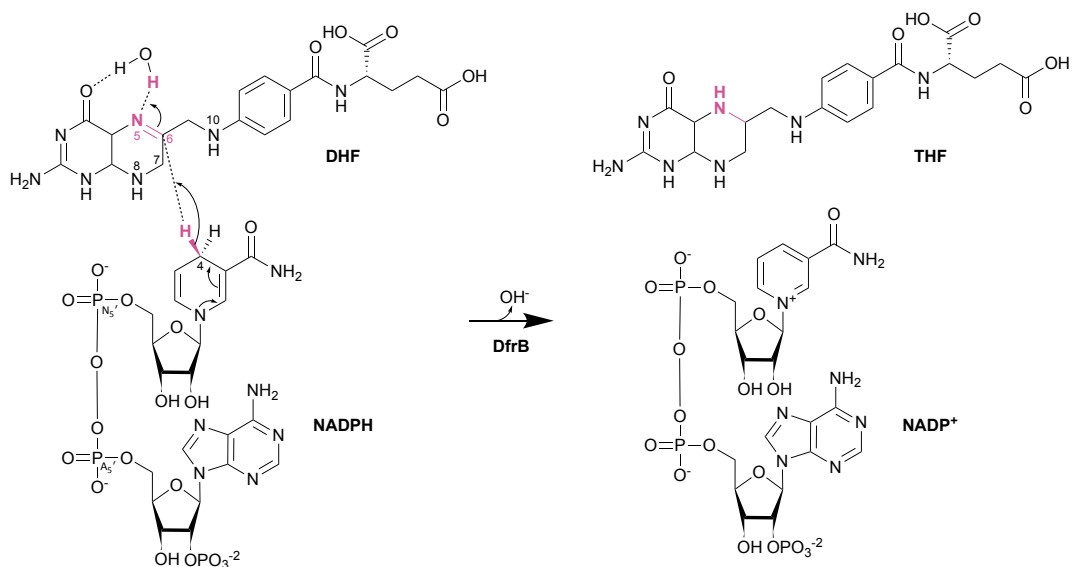
Enzymes are complex macromolecules capable of catalyzing a wide variety of chemical reactions with high efficiency. Nevertheless, biological catalysis can be rudimentary. Here, we describe an enzyme that is built from a simple protein fold. This short protein sequence – almost a peptide – belongs to the ancient SH3 family of binding modules. Surprisingly, this binding module catalyzes the specific reduction of dihydrofolate using NADPH as a reducing cofactor, making this a dihydrofolate reductase. Too small to provide all the required binding and catalytic machinery on its own, it homotetramerizes, thus creating a large, central active site environment. Remarkably, none of the active site residues are essential to the catalytic function. Instead, backbone interactions juxtapose the reducing cofactor proximal to the target imine of the folate substrate, and a specific motion of the substrate promotes formation of the transition state. In this feature article, we describe the features that make this small protein a functional enzyme capable of catalyzing a metabolically essential reaction, highlighting the characteristics that make it a model for the evolution of primitive enzymes from binding modules.

2.2 Introduction

Enzymes are essential to life, increasing the rate of chemical reactions in solution by up to 10^{18} -fold compared to the uncatalyzed reactions.¹ They accelerate chemical reactions with an efficiency that can be so high that it is limited only by the diffusion of molecules in the solvent, with catalytic efficiencies reaching $10^{10} \text{ M}^{-1} \text{ s}^{-1}$.²⁻⁵ The vast array of reactions catalyzed by enzymes may seem at odds with the modest chemical diversity of the functional groups that comprise them. How does a simple protein provide life-sustaining catalytic activity?

Diverse catalytic strategies have naturally evolved to enable enzymes to accelerate the rate of chemical reactions. These include transition-state stabilization, substrate destabilization, general acid-base catalysis and the formation of covalent enzyme-substrate intermediates.⁶ Whereas some of these strategies are mechanistically complex, others are straightforward. In particular, proximity-based catalysis relies on aligning the reactive chemical groups with the optimal geometry. Proximity-based catalysis increases the local concentration of the reagents, favoring catalysis with no further contribution to the chemical transformation by the enzyme.⁷

Type B dihydrofolate reductase (DfrB) enzymes exploit such a catalytic strategy to accelerate the production of the essential metabolite tetrahydrofolate (THF; Scheme 2.1). Essentially, the active site serves as an environment to juxtapose and orient the conjugated but non-aromatic N5=C6 imine of dihydrofolate (DHF) with the hydride contributed by β -nicotinamide adenine dinucleotide phosphate (NADPH).^{8,9}



Scheme 2.1. Chemical reaction catalyzed by the DfrB enzymes.

Following protonation of the N5 of dihydrofolate (DHF) by the solvent, the NADPH hydride is transferred to the C6 of DHF. The product, tetrahydrofolate (THF), serves as an essential carrier of ‘one-carbon units’ in core metabolic reactions that include purine biosynthesis.

The mechanism of the reduction reaction appears to be fundamentally flawed for several reasons that will be discussed below. An illustration of its apparently suboptimal (natural!) design is that the hydride transfer requires preprotonation of the DHF imine by the solvent rather than by the enzyme.¹⁰ This is inefficient because the calculated pK_a of the DHF N5=C6 imine is 4.5, which is 2-3 orders of magnitude below the pH of the microbial cytoplasm where the reaction occurs.¹¹⁻¹³

As a result of this and other atypical properties (*vide infra*), DfrB enzymes have been described as ‘primitive’.¹⁴ Nevertheless, they are functional under biological conditions of substrate availability, as is generally observed for enzymes.⁴ Moreover, their seemingly mediocre design is counterbalanced by a key, unexpected feature: DfrB enzymes confer insurmountable resistance to the broadly prescribed and highly effective synthetic antibiotic, trimethoprim (TMP) – a property that has given them an edge, promoting their spread in pathogenic organisms.^{15,16}

The DfrB enzymes are structurally and evolutionarily unrelated to the better known ubiquitous dihydrofolate reductases (FolA enzymes). FolA provides metabolically essential THF to almost all living cells. Although the microbial and mammalian FolA are evolutionarily related, structural differences between them can be leveraged to specifically inhibit the microbial FolA using TMP. However, as a result of their evolutionarily distinct structures, DfrB enzymes are not inhibited by TMP or by most antifolate-

type antibiotics that inhibit the microbial Foa family.^{15,17} Thus, the intrinsic resistance of the DfrB to the antimicrobial TMP has long been thought to have promoted their clinical emergence.^{15,16}

The prototypical DfrB1 enzyme was first identified in pathogenic bacteria in the 1970s in the context of clinical resistance to TMP, allowing for its early characterization.^{18,19} DfrB1 (formerly known as R67 DHFR or DfrII) was found to be little more than a peptide. In fact, only 56 of its 78 residues are required for activity, making it one of the shortest catalytic protein sequences known.²⁰ Apart from catalyzing the same reaction, the Foa and DfrB enzyme families share no common features.¹⁴

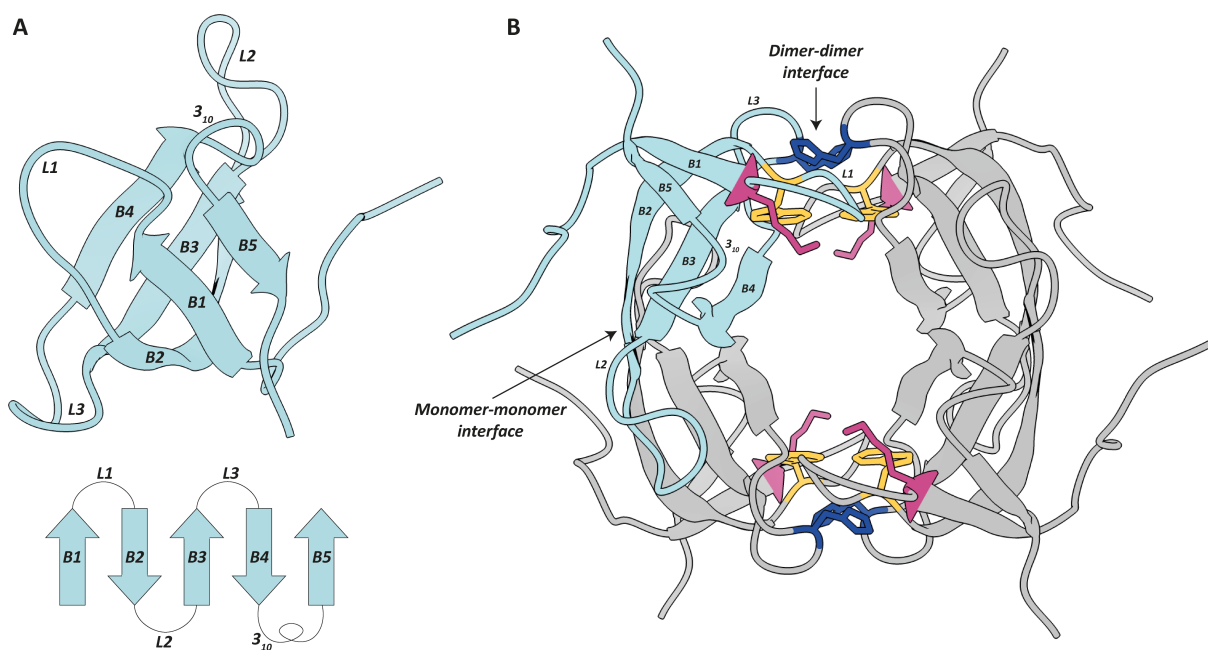


Figure 2.1. The crystal structure of the DfrB1 enzyme (2rk1 PDB).

A. Topology of the SH3 fold of each DfrB1 protomer. **B.** Homotetrameric structure of the functional DfrB enzyme. One protomer is colored blue while the three identical protomers are colored grey. The symmetrical dimer-dimer interfaces are at the top and bottom of the tetramer, where the symmetrical Lys32 (fuchsia), Trp38 (gold) and His62 (dark blue) from each of the four protomers are represented in sticks.

How can a 56-amino acid protein domain effectively bind and orient two bulky molecules (Scheme 2.1) to catalyze hydride transfer at a metabolically relevant rate? To do this, the DfrB1 protomer self-assembles into its active, homotetrameric form; all four protomers form the enzyme's single, central active site – an effective strategy for amplifying the binding surface area (Figure 2.1). The homotetrameric structure of the DfrB active site is particularly unusual. It is extremely rare in nature for the single active site of a

homomultimeric enzyme to be formed by more than one monomer, the only other example we know of being the homodimeric HIV-1 protease.^{21,22}

In this feature article, we review the evidence accumulated over the past five decades to argue that DfrB enzymes originated from a small protein domain that served as a broad-purpose binding module. Among its functions, it bound nucleotides such as the dinucleotide NADPH and other metabolites such as DHF. An evolutionary iteration that bound both NADPH and DHF allowed the metabolically-advantageous catalytic transformation of DHF to THF. And with the introduction of the cheap and effective antibiotic TMP to clinical and veterinary settings worldwide in the 1960s, the stage was set for pathogenic microbes to accelerate the dissemination of this ‘primitive’ yet highly TMP-resistant source of life-sustaining THF.

2.3 The first step to catalysis: creation of a binding scaffold

The 78-residue DfrB1 protein consists of an intrinsically disordered, 18-residue *N*-terminal region and a 60-residue SRC homology 3 (SH3) domain.²³ The SH3 fold is one of the most ancient and abundant protein modules in nature.^{24–26} Widely included in larger multi-domain proteins, SH3 domains are generally responsible for mediating protein-protein interactions by binding proline-rich peptide motifs.²⁷ SH3 domains have also evolved to bind to a wide variety of metabolites such as nucleotides, further illustrating their great binding promiscuity.²⁸ The DfrB1 SH3 domain is a β -barrel composed of five antiparallel β -strands connected by three short loops and one turn composed of a short 3_{10} helix (Figure 2.1A).²³

The SH3 domain of DfrB1 self-assembles, first into a dimer and then into a tetramer, forming two distinct types of interfaces.^{29,30} To dimerize, residues from the B2 strand of two protomers H-bond, such that the B2 to B4 strands of both protomers form an extended β -sheet (Figure 2.1B). Although only three residues from each protomer contribute to this intersubunit core, the monomeric species is rare, with $K_D(\text{monomer-dimer}) = 50 \text{ pM}$.³¹

The active form of the enzyme is the homotetramer; the dimeric form shows no significant activity.³² The tetramer is formed upon dimerization of two dimers, through reciprocal loop interactions between L1 of each dimer and L3 of the dimer it faces (Figure 2.1B). The tetrameric assembly is 2-3 orders of magnitude weaker than the dimer assembly, with $K_D(\text{dimer-tetramer}) = 10 - 50 \text{ }\mu\text{M}$.^{33,34} Each of the two identical dimer-dimer interfaces contains a pair of histidines (one His62 from each L3) that lie within $\sim 4 \text{ \AA}$ of each other (Figure 2.1B). At neutral pH (7-8), the active homotetramer predominates due to H-bonding between the neutral histidines;^{30,33} at pH 5, the protein is predominantly dimeric – and therefore largely inactive – due to repulsion between the protonated histidines ($\text{p}K_2^{\text{His}} 5.97$).³⁵ This observation confirmed that the four His62 are major contributors to tetramer stability.³⁰ Indeed, substitution of His62 to cysteine obviates this

pH dependence and instead promotes the formation of a disulfide bridge that stabilizes the dimer-dimer interface.¹⁰

At the same interface, a tryptophan in L1 (Trp38) also contributes to tetramerization (Figure 2.1B). Its substitution with phenylalanine destabilizes the tetramer into a predominantly dimeric population, thereby reducing the catalytic efficiency by 10^3 -fold.^{32,36} Also at the dimer-dimer interface, chemical modification of His62 or, alternatively, the substitution of Lys32 with alanine or methionine, destabilize the tetramer and shift the population toward dimers.^{33,37} Interestingly, addition of one of the reaction substrates, NADPH, stabilizes the tetrameric population. This highlights the importance of protein-ligand interactions in promoting the formation of the active tetramer.^{36,38}

The tetrameric complex is rigid, with little backbone motion observed either by NMR or by molecular dynamics (MD) simulations.^{9,39} This rigidity is accompanied by high thermal tolerance: a melting event $\geq 60^\circ\text{C}$ appears to indicate tetramer dissociation since secondary structure is largely maintained.^{40,41} Melting is reversible, allowing $>90\%$ recovery of catalytic activity after 10 min at 95°C .^{36,42,43}

The SH3 core is highly conserved within the 20-member family of DfrB enzymes, highlighting the essential nature of tetramer formation for activity (Figure 2.2). In contrast, the only conserved feature of the ~ 20 -residue, disordered *N*-terminal region of the DfrB protomer is its length.^{15,23,44} Consistent with this, the creation of all possible point substitutions in the enzyme through a deep mutational scanning approach has demonstrated that substitutions in the SH3 core of DfrB1 are detrimental to the enzyme fitness, whereas the *N*-terminal region is largely tolerant to sequence alteration.⁴⁵ Small-angle neutron scattering, MD simulations and ^{19}F NMR have determined that the *N*-terminal region adopts a globally compact shape and interacts with Trp45 at the monomer-monomer interface, providing 2.6 kcal/mol of stability to the assembly.^{40,46} Although deletion of the *N*-terminal encoding region from the DfrB1 gene results in an active, TMP-resistant enzyme, the *N*-terminal region has been shown to contribute to the formation of the active form of the enzyme. For example, certain DfrB variants that are inactive when only the SH3 fold is expressed are active when expressed following the *N*-terminal region, suggesting a role in stabilizing the structure or contributing to folding.²⁰ Truncation of the *N*-terminal 16 residues after expression also results in a fully active enzyme that is functionally indistinguishable from the full-length enzyme.^{16,20,31}

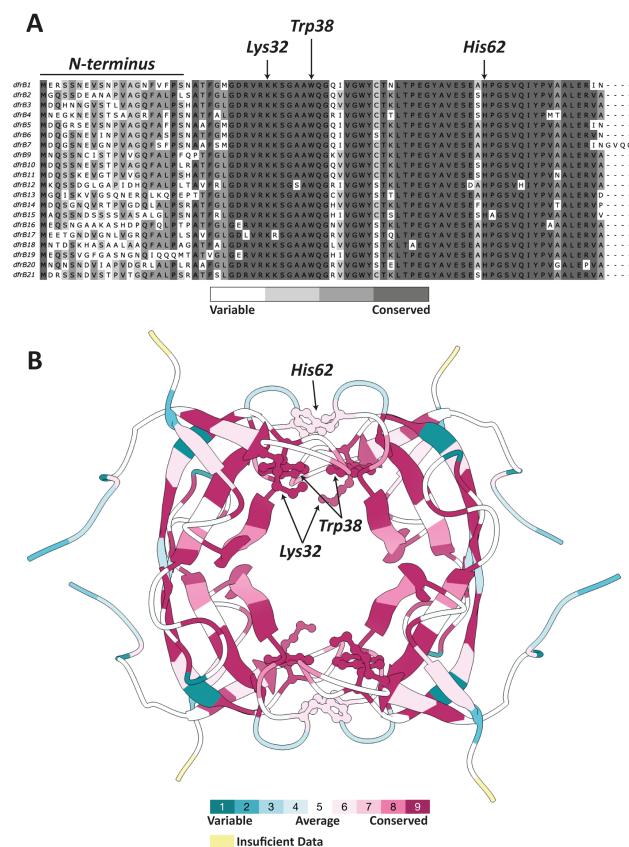


Figure 2.2. Conservation of the DfrB1 sequence.

A. Multiple sequence alignment of the 20 characterized members of the DfrB family generated by MAFFT and displayed with UGENE.¹⁰⁸ Conserved residues are highlighted in dark grey whereas variable residues are highlighted in white. **B.** Sequence conservation of the DfrB domain mapped on the structure (2rk1 PDB) using ConSurf.¹⁰⁹ Highly conserved residues are colored dark pink, whereas variable residues are colored dark blue. Residues Lys32, Trp38 and His62, important contributors to the dimer-dimer interface, are represented in ball and sticks.

2.4 From scaffold to catalyst: positioning the reagent proximal to a reducing agent

The homotetrameric complex creates a single, central tunnel to which the four protomers contribute. Remarkably, the tunnel has a tetrahedral symmetry described by three 2-fold rotation axes. Thus, every position in the pore has three other equivalent positions.

This highly symmetrical cavity has evolved from the binding function that characterizes SH3 domains, into a catalyst: the hydride-donating NADPH enters by one of the two identical tunnel mouths and the DHF substrate enters by the opposite mouth to meet it in the middle, where the reduction reaction happens (Figure 2.3.).

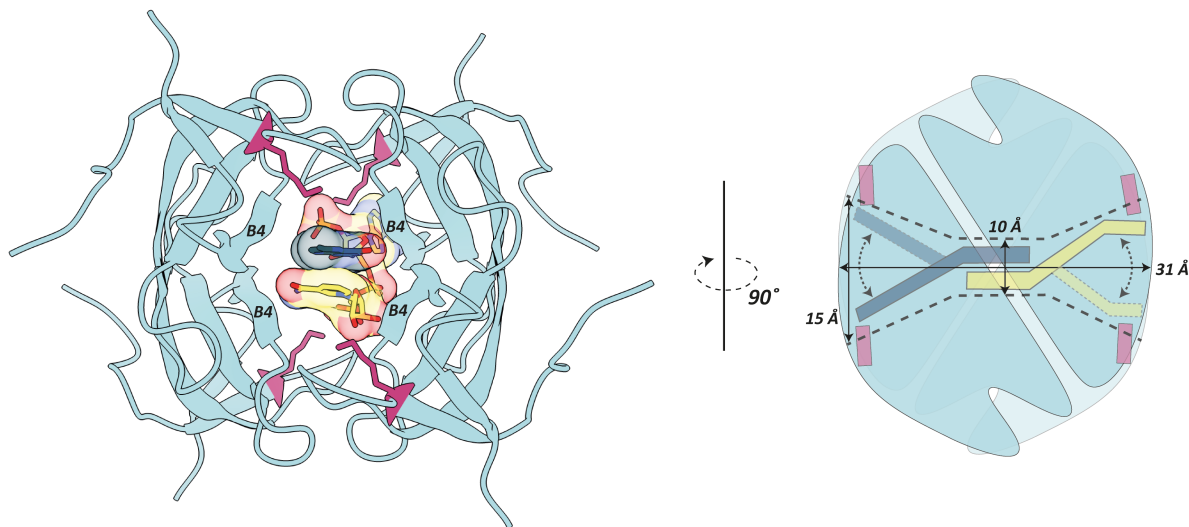


Figure 2.3. Complex of the DfrB1 enzyme with the cofactor NADP⁺ (yellow) and the pterin of DHF (dark blue) (2rk1 PDB).

Left: The four Lys32, each from one protomer, are represented in pink sticks. The tail of DHF is not resolved in the crystal structure. Right: schematic representation of the complex. The active-site tunnel is outlined with dashed lines, with its dimensions indicated. The movement of the ‘tails’ of DHF and NADP⁺ between symmetrical Lys32 at either tunnel mouth is represented in dotted curved arrows.

This immediately raises a flag. Biocatalysts typically procure highly specific binding surfaces for their ligands, particularly those that catalyze the synthesis of essential metabolites.⁴⁷ The FolA-type dihydrofolate reductase enzymes exemplify this: the narrow active-site cleft is highly complementary to DHF and to NADPH, each binding at a precise site and forming multiple specific contacts with the enzyme.⁴⁸ In contrast, each of the four ‘inner walls’ of the DfrB tunnel is formed by an identical B4 strand (Figure 2.3.). It follows that the inner binding surfaces and the tunnel mouths are not specifically tailored to binding of either DHF or NADPH, yet they can bind both. Indeed, the crystal structure reveals that the ligands bind by mediating different interactions with symmetry-related residues.⁸ Thus, the primary function of the SH3 domain, which is to create protein-protein interactions, has been co-opted into a catalytic function.

Each identical entrance to the active site houses two symmetry-related Lys32 residues, that create two ionic interactions with the negatively charged P_{N5'} and P_{A5'} of NADPH (Scheme 2.1).⁸ This was confirmed by a 50-fold drop in catalytic efficiency when using NADH as a cofactor.³⁷ The Lys32 also bind the negatively charged *p*-aminobenzoyl glutamate (*p*ABA-Glu) tail of the DHF substrate.^{8,23,49} The contribution of this interaction is made evident by the contrast with dihydrobiopterin, a DHF analog lacking the *p*ABA-Glu tail, that shows no binding signal by isothermal titration calorimetry (ITC).⁵⁰ The reduction of

dihydropteroate is significantly slower than that of DHF, with a ~ 10 -fold increase in K_M and a ~ 1600 -fold reduction in k_{cat} .⁵¹

Despite forming ionic interactions with Lys32, the DHF *p*ABA-Glu tail cannot be resolved experimentally, suggesting high flexibility.¹⁵ MD simulations have confirmed that the tails of both DHF and NADPH exhibit large fluctuations, where their carboxylate and phosphate groups, respectively, interact alternately with both Lys32 at each tunnel entrance (Figure 2.3.). This high flexibility (or ‘wagging’) of the ligand tails occurs across each of the 15 Å-wide, funnel-like tunnel entries that each hold two Lys32 on opposite faces.^{9,49} Indeed, the hourglass-shaped tunnel is voluminous: 31 Å long and narrowing from a width of 15 Å at the opening to ~ 10 Å at its center, it can hold up to 168 water molecules.⁵²

The four ‘walls’ of the active site are formed by the four-fold B4 strand (Figure 2.4). Its Val66-Gln67-Ile68-Tyr69 (VQIY) motif is highly conserved within the DfrB family (Figure 2.2A).⁴⁴ Both above and below the center of the tunnel, two pairs of Gln67 and Tyr69 form a hydrogen bonding network that shapes the narrower tunnel centre.⁸

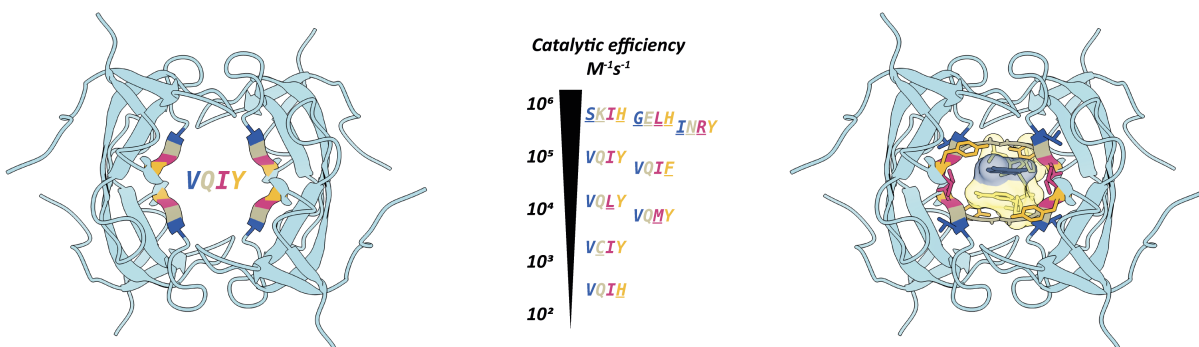
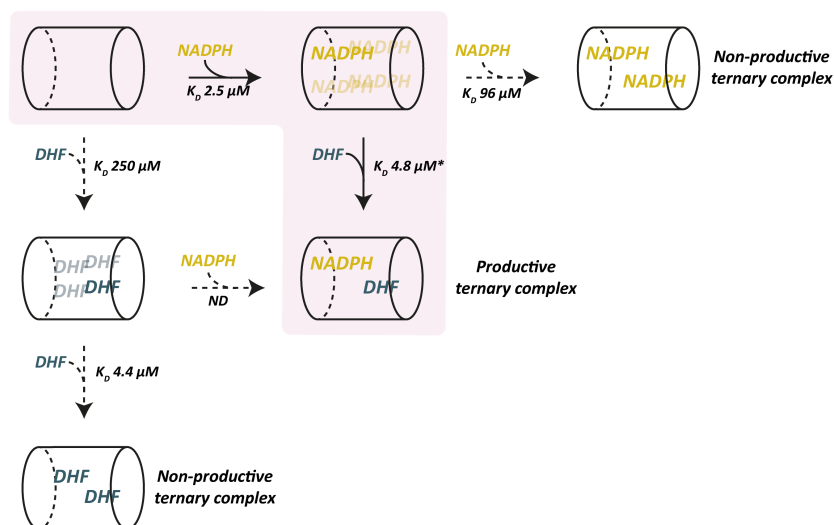


Figure 2.4. The VQIY active-site motif and its variations in engineered DfrB1 variants.

The active-site Val66, Gln67, Ile68 and Tyr69 are respectively colored blue, beige, fuchsia and gold. Alterations to the active-site motif in engineered variants and their catalytic efficiency are indicated in the centre, with substituted residues underlined. Results are compiled from the following references.^{54,69,86} As represented on the right, the four Tyr69 side-chains form a clamp that defines the tunnel width, yet all interactions with the pterin ring of DHF (dark blue) and the nicotinamide ring (yellow) are established with the protein backbone.

NADPH is the first to bind to the enzyme, with a K_D of 2.5 μM .⁵⁰ This promotes binding of DHF (K_D 4.8 μM for the DfrB1-NADP⁺ complex), forming the catalytic ternary complex (Scheme 2.2).⁵³ Indeed, binding of DHF in the absence of NADPH is unlikely (K_D 250 μM).⁵³ As a consequence of the four-fold symmetry, a second NADPH can bind to the DfrB1-NADPH complex albeit with negative cooperativity, with a K_D of 96 μM .⁵⁰ It is even possible for two molecules of DHF to bind simultaneously in the active site: following the improbable event of DHF binding first, a second DHF-binding event is promoted (K_D 4.4 μM).⁵⁰ This

potential for non-productive binding of NADPH-NADPH or DHF-DHF highlight the ‘primitive’ nature of this protein assembly, detailed below.²³



Scheme 2.2. Ligand binding to DfrB.

The active-site tunnel is represented as a hollow cylinder; it has two identical mouths. Due to the symmetry of the tunnel, there are four identical binding surfaces; the first molecule to enter in the pore can bind to any of the four sites, which are represented in lighter lettering. The favored path to formation of the productive ternary complex is highlighted. *Binding of DHF to form the ternary complex was determined with bound NADP⁺. ND = not determined. Results are compiled from the following references.^{50,53}

This positive, inter-ligand cooperativity is the critical feature that makes DfrB an enzyme rather than a binding module; without the formation of the DfrB-NADPH-DHF ternary complex being favored, no reaction could occur.^{54,55} The source of the positive cooperativity may, at first, seem perplexing: no protein conformational shifts are observed when comparing the apo enzyme and the productive ternary complex, suggesting that positive cooperativity is due to inter-ligand interactions rather than protein structure rearrangement.⁵⁵ The key is the initial binding event. Once NADPH is bound, the active-site symmetry of DfrB is broken; DHF then stacks onto the bound NADPH within the tunnel.

Active-site substitution of the VQIY motif to VHIY shifts the DfrB from a catalytic protein back toward a binding module. The substitution of Gln67 to His enhances the binding of NADPH and DHF to the apo enzyme ($K_D^{\text{NADPH}} 0.054 \mu\text{M}$ and $K_D^{\text{DHF}} 0.040 \mu\text{M}$).^{55,56} However, binding cooperativity is lost and catalytic turnover rate drops nearly two orders of magnitude, from 1.3 s^{-1} to 0.025 s^{-1} .⁵⁵ This suggests that Gln67 plays an important role in promoting formation of the productive ternary complex.

2.5 The proximity-based catalytic mechanism of DfrB

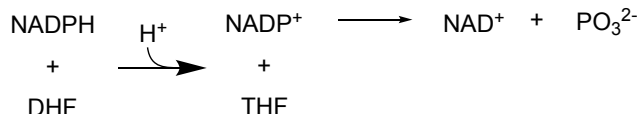
The dihydrofolate reductase reaction consists of hydride transfer from NADPH to DHF-C6, accompanied by protonation of DHF-N5. The chemistry illustrated in Scheme 1 applies to both DfrB and to Fola. However, the reaction mechanism of DfrB is two orders of magnitude less efficient than that of Fola, as illustrated both by the hydride transfer rate of 1.3 s^{-1} for DfrB relative to 238 s^{-1} for *E. coli* Fola and their respective catalytic efficiencies of $2.2 \times 10^5 \text{ M}^{-1}\text{s}^{-1}$ and $2.4 \times 10^7 \text{ M}^{-1}\text{s}^{-1}$.^{10,43,57}

One factor accounting for this difference is the protonation of DHF-N5. In *E. coli* Fola, the protonation step is efficient. Active-site residues Asp27 and Tyr100 increase the $\text{p}K_{\text{a}}$ of DHF-N5 from 2.6 to 6.7, facilitating its solvent-mediated protonation.⁵⁸⁻⁶¹ In contrast, the absence of a general acid in the solvent-accessible active site of DfrB is one of the reasons why it is described as unevolved; the $\text{p}K_{\text{a}}$ of DHF-N5 is calculated to be 4.5 in DfrB1.¹¹⁻¹³ Nevertheless, DfrB1 is inactive at low pH where protonation of the four His62 prevents tetramer formation (Figure 2.2). Using the His62Cys variant which remains tetrameric at pH 5, the turnover rate was found to be proportional to the solvent proton concentration, with a k_{cat} up to $150,000 \text{ s}^{-1}$ at pH 4.95.¹⁰ This demonstrates that higher catalytic efficiency could have been achieved in nature. That DfrB1 did not evolve to have high efficiency suggests that the pH-dependency of tetramer formation is an evolutionary advantage, or yet that a higher efficiency does not confer a survival advantage in the native context.

In *E. coli* Fola, the rate-limiting step for catalysis is product release.^{62,63} In contrast, primary isotope effect studies of DfrB1 indicate that hydride transfer is at least partially rate-determining.⁴³ Leading up to hydride transfer, NADPH and the pterin ring of DHF are rigidly bound at the centre of the tunnel. The stacking of the nicotinamide and the pterin rings stabilizes the *endo* transition state.^{8,49} Whereas the binding of NADPH and DHF to DfrB is enthalpy-driven, a large entropic component is associated with catalysis.⁵¹ In fact, motions of Gln67, near C₄ of NADPH (Scheme 2.1), favor ring puckering, which promotes formation of the transition state. Puckering of the DHF pterin ring results from the large movements sampled by the carboxylate groups at the tail of DHF, as they oscillate between the symmetry-related Lys32 at the tunnel mouth.^{9,49} Thus, although counter-intuitive, ‘wagging’ of the DHF tail is needed to reach the transition state, resulting in *substrate disorder-assisted* catalysis.⁶⁴ Indeed, substitution of Lys32 results in loss of ion pairing and thus reduced TS stabilization, correlating with a reduced k_{cat} .^{49,65}

Although the *pro*-R hydrogen of NADPH is transferred to the DHF-C6 in both types of dihydrofolate reductase, the ring-stacked *endo* transition state of DfrB is distinct from that of *E. coli* Fola.^{8,66} In Fola, the *exo* transition state is stabilized, with minimal overlap between the nicotinamide and pterin rings.⁶⁷ The distinct approach to the transition state further illustrates how these evolutionarily unrelated enzymes differ.^{14,54}

Interestingly, kinetic characterization of DfrB1 led to the discovery of its capacity to slowly dephosphorylate the reaction side-product NADP^+ , yielding NAD^+ (Scheme 2.3).^{39,68} This weak phosphate hydrolase activity (turnover of $1.1 \times 10^{-5} \text{ s}^{-1}$) highlights the capacity of the DfrB1 to create an environment conducive of catalysis that is not exclusive to dihydrofolate reduction. Perhaps hydrolase reactivity of DfrB with structurally related molecules is yet to be discovered.



Scheme 2.3. Dihydrofolate reductase and NADP^+ phosphatase reactions catalyzed by DfrB1.

2.6 A permissive active site: catalysis is maintained with modifications to either the active site of the cofactor

Point substitutions to the conserved VQIY motif on the B4 strand, such as VCIIY and VQIH, reduce the catalytic efficiency of DfrB1 either by increasing K_M of the substrates or by decreasing k_{cat} (Figure 2.4).⁵⁴ We note that each point mutation in the gene results in four identical active-site substitutions in the enzyme, magnifying their impact on catalysis. However, our screening of highly substituted active-site variants brought to light the high tolerance of DfrB1 to concerted active-site alteration. The simultaneous substitution of three or even four residues of the VQIY motif (to SKIH, INRY, GELH) yielded enzymes with 3- to 6-fold increased catalytic efficiency for dihydrofolate reduction (Figure 2.4).⁶⁹

Importantly, most of the highly modified active-site motifs were inactive, only 0.2% of the explored motifs yielding a functional enzyme. Variants could potentially lack activity due to formation of an active site not conducive to catalysis, alteration of the SH3 fold, inability to tetramerize, protein instability, or lack of expression. The tolerance to active site substitutions is consistent with the productive juxtaposition of the nicotinamide and pterin rings being primarily mediated by the backbone of active-site residues, rather than by their side-chains. Hence, substitutions yielding backbone geometry compatible with ligand stacking to achieve transition-state stabilization can yield an active enzyme, regardless of the side-chains. Nonetheless, no pattern identifying physical or chemical properties essential to transition-state stabilization has been identified from the collection of functional motifs identified to date (Figure 2.4).⁶⁹

Both FoaA and DfrB1 enzymes show weak reactivity with naturally-occurring NADH (which lacks the 2'-phosphate), demonstrating a certain permissivity regarding the cofactor accepted for dihydrofolate

reduction.⁷⁰ However, DfrB1 is more tolerant than Fola to non-naturally occurring reducing cofactors. Only DfrB1 can use the non-natural α -NADPH (carrying the α -anomer of the nicotinamide rather than the naturally occurring β -anomer), with a 4-fold increase in K_M , or thio-NADPH (where the nicotinamide amide oxygen is substituted by sulfur), with a 2-fold increase in K_M .⁷⁰ In contrast, the chromosomal dihydrofolate reductases (Fola) from bacteria, mammals and plants are incapable of reducing dihydrofolate with either of these compounds, as no activity could be detected even when the enzyme concentration was increased by 5,000 to 10,000-fold.⁷⁰ Although there exists no evolutionary pressure to exclude reactivity with compounds that are not naturally-occurring, this highlights the broader tolerance of DfrB1 to substitutions in the dinucleotide cofactor, and is a further example of the evolutionary divergence of the DfrB and Fola.

2.7 Learning more about binding by inhibiting: discovery of DfrB inhibitors

The symmetry of the voluminous active site of DfrB1 allows for different combinations of ligands to bind in the tunnel.⁵⁰ While the binding of DHF and NADPH is the favored and productive combination of ligands, the binding of two copies of DHF or NADPH is also possible (Scheme 2.2). We exploited this symmetrical binding potential to design DfrB inhibitors.

The active site of the DfrB being structurally and evolutionary unrelated to the ones of Fola, inhibitors of these latter enzymes – such as TMP, methotrexate and aminopterin – are not relevant as inhibitors of the DfrB family.¹⁷ Our first strategy in DfrB inhibitor discovery was to perform fragment-based inhibitor design inspired by the structures of DHF and NADPH.⁷¹ We identified small molecules that weakly yet selectively inhibit DfrB1, sparing the human dihydrofolate reductase. From molecules inhibiting DfrB1 activity in the micromolar range, we designed symmetrical bis-benzimidazole-based molecules bearing terminal carboxylates to bind Lys32 at the mouths of the active site. This first series of low micromolar inhibitors binds in two copies in the active site, providing competitive inhibition (Figure 2.5).⁷¹ Further structure-activity relationship (SAR) studies confirmed the contribution of the benzimidazoles to binding, and the role of inhibitor length in establishment of essential interactions of the terminal carboxylates with the Lys32.⁷² In an attempt to reduce the entropic cost of binding two inhibitor molecules, we designed a tetra-benzimidazole analog, as well as a V-shaped analogue based on an *ortho*-substituted central phenyl core. As per their design, they bind as a single copy to the DfrB active site. Nonetheless, their binding constants were not improved.⁷²

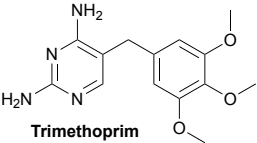
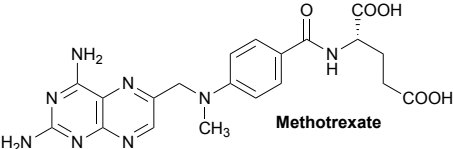
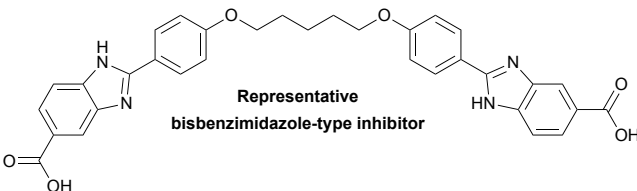
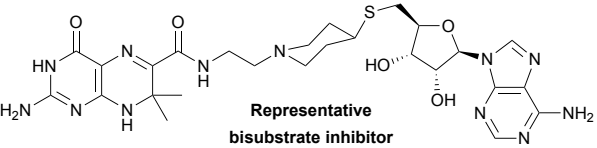
	K_i (μM)	
	DfrB	<i>E. coli</i> FoaA
 <p>Trimethoprim</p>	150 ^a	$(20 \pm 7) \times 10^{-6}$ ^b
 <p>Methotrexate</p>	Cannot be determined ^c	$(1.0 \pm 0.6) \times 10^{-6}$ ^d
 <p>Representative bisbenzimidazole-type inhibitor</p>	2.0 ± 0.3 ^e	ND
 <p>Representative bisubstrate inhibitor</p>	20 ± 3 ^f	ND

Figure 2.5. Inhibitors of the unrelated dihydrofolate reductase enzymes DfrB and FoaA.

Inhibitors specific to *E. coli* FoaA do not inhibit DfrB enzymes efficiently. The DfrB K_i for trimethoprim was reported for DfrB2, and is representative of all other DfrB enzymes assayed. The K_i for the bisbenzimidazole-type and bisubstrate inhibitors were determined for DfrB1. ND = not determined. Reference ^{a 110, b 111, c 17, d 112, e 71, f 42}.

We further identified a second class of DfrB inhibitors. Initially designed as inhibitors of the 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK), another enzyme belonging to the folate pathway, these inhibitors are based on the pterin moiety of DHF and the adenosine of NADH and are therefore named ‘bisubstrate’ (Figure 2.5).⁴² Upon improvement by SAR, these bisubstrate, dual-target inhibitors inhibit the DfrB in the micromolar range. Interestingly, they appear to bind in the DfrB tunnel in complex with NADPH.⁴²

These are the two first classes of selective inhibitors reported for DfrB enzymes. We have found them to be promising tools for investigation of novel DfrB sequences, as mentioned below. Interestingly, inhibitors of both classes successfully inhibit all DfrB members against which they were assayed.⁴² We are currently establishing high-throughput assays to characterize kinetic parameters, TMP resistance as well as inhibition, to expand our knowledge of this emerging family of resistance enzymes.

2.8 Poorly evolved catalysts: the DfrB family exhibit characteristics of primitive enzymes

It has been proposed that the evolution of highly efficient and specific modern enzymes over the past four billion years began with promiscuous, multifunctional precursors.^{73,74} These primitive enzymes were sufficient to perform all the metabolic reactions necessary for the survival and proliferation of primordial cells. Duplication and selection of these rudimentary enzymes led to their specialization.⁷⁵

The functions of the first enzymes are thought to be related to phosphate binding and nucleotide metabolism is proposed to rely on the most ancient network of enzymes.^{76–79} The cystathionine β -lyase enzymes of *Pelagibacter ubique*, *Drosophila melanogaster* and *Thermotoga maritima*, which also catalyze the unrelated alanine racemase activity, are examples of such primitive enzymes.⁸⁰ These multifunctional enzymes necessarily had biologically-relevant K_M values in order for substrate binding to occur; however, they turned over slowly, their poor k_{cat} reflecting their lack of specialization. Despite their poor catalytic efficiency, these enzymes provided sufficient activity to support host proliferation.

The DfrB enzymes share many of the characteristics of primordial enzymes, as summarized below:

2.8.1 The DfrB enzymes are not specifically evolved for efficient dihydrofolate reduction

Primordial enzymes are described for their ability to catalyze a variety of chemical transformations, allowing the survival of organisms with a limited set of proteins. The weak NADP⁺ hydrolase activity of DfrB1 may be indicative of such reaction promiscuity.³⁹ While the K_M of primordial enzymes must be similar to those of modern enzymes to be matched to cellular substrate concentrations, their k_{cat} is often weaker than those of modern enzymes.^{80,81} This is the case of the DfrB, with a K_M in the metabolically-relevant micromolar range and $k_{cat} < 0.5 \text{ s}^{-1}$, which is 20-fold slower than Fola.⁴² Nevertheless, this rate is adequate to maintain sufficient quantities of THF in the cell while Fola is inhibited by TMP, thus providing trimethoprim resistance in the pathogenic bacteria where DfrB were first discovered.¹⁹

The proximity-based mechanism of the DfrB for dihydrofolate reduction is simple: positive cooperativity of substrate binding allows entry and correct positioning within the pore.⁵⁴ Proximity of the DHF pterin C6 and the NADPH nicotinamide assisted by disorder of the DHF tail promotes hydride transfer; however, proton uptake from the solvent by the relatively acidic N5=C6 imine is not facilitated. In contrast, the ubiquitous and highly efficient Fola dihydrofolate reductases have evolved an acidic residue to protonate the N5=C6 imine, thereby accelerating hydride transfer to C6.¹⁴ The lack of an active-site general acid in DfrB justifies the two orders of magnitude slower catalytic efficiency.⁴²

2.8.2 A highly symmetrical pore is a poor design for an active site

The shape of the central pore formed by homotetramerization is an important indicator of the poorly evolved nature of this enzyme. Due to the 222 symmetry of the active site, four-fold cumulative effects accompany any sequence alteration. This unusual feature severely curtails the evolutionary potential for mutations that could improve the catalytic mechanism or specificity. Indeed, the four-fold symmetrical residues form different interactions with DHF and NADPH. Therefore, alteration of the active site to favor the binding of one molecule is likely to disfavor binding of the other.

Duplication of the DfrB gene for translation in tandem as a single polypeptide could theoretically overcome this limitation. Indeed, the strategy of duplication in homomeric complexes is observed in other proteins,^{82,83} and laboratory creation of tandem DfrB, involving up to four, fused repeats, yielded a functional enzyme with similar efficiency.^{84,85} The tandem DfrB constructs allowed examination of asymmetric substitutions; despite informing on the importance of specific interactions between residues and substrates, none yielded a more efficient enzyme.^{41,55,65,86} Importantly, natural chain duplication has never been observed for DfrB enzymes, suggesting that such a duplication does not provide a clear evolutionary advantage.

In addition to the fact that ligand entry into either of the two identical tunnel mouths is equally probable, the tunnel is unusually voluminous – over 3600\AA^3 , which is three times larger than average.^{14,87} Consequently, the tunnel holds approximately 168 water molecules, greatly surpassing what is characteristic of modern enzymes.⁵² Interestingly, asymmetric substitutions that reduce the volume of the half-pore by about 35% had no significant impact on catalysis, demonstrating that the catalytic activity is resistant to important morphological changes in the active site pore.⁶⁴ This provides further evidence that the DfrB enzyme is poorly evolved for dihydrofolate reduction.

Nevertheless, the active-site symmetry is advantageously exploited. The negatively charged tails of DHF and of NADPH explore both equivalent binding surfaces at either tunnel mouths, alternating between the symmetrical Lys32 for binding. This symmetry-induced disorder is essential to the catalytic mechanism.

2.8.3 The DfrB family lacks the specialized properties that are typical of modern enzymes

A particularity of DfrB1 is that none of its active site residues are indispensable for catalysis. Activity can even be maintained upon simultaneous substitution of all four residues of the VQIY motif (Figure 2.4).⁶⁹ This highlights the fact that no amino acid side-chain is essential for catalysis. Instead, the DfrB create an environment shaped by side-chains, where the protein backbone is the main contributor to ligand binding, a nucleotide binding mode considered to be ancient.⁸⁸ The productive juxtaposition of DHF and NADPH at the center of the pore can thus be mediated by other residue combinations.⁸ A factor constraining the

identity of the active site residues, located on the B4 strand, is their need to preserve the typical β -barrel structure describing the SH3 fold.

Thus, the driving force of catalysis is not specific to side-chain chemistry, but rather to interactions between the ligands. Although simultaneous binding of two molecules of DHF or of NADPH in the pore is possible, positive cooperativity between the ligands favors the formation of the productive DfrB1-NADPH-DHF ternary complex (Scheme 2.2).⁵⁵

We hypothesize that the evolution of dihydrofolate reductase activity from this SH3 fold occurred in a stepwise fashion. At least two orders of evolutionary events can be evoked: multimerization followed by substrate binding within the pore, or substrate binding to the SH3 fold followed by multimerization. Since ancient SH3 domains are known to bind nucleic acids, the similarity of pterin-based metabolites to nucleotides may have facilitated the cooperative binding of the dinucleotide NADPH and of DHF in the pore.²⁶ Importantly, THF and NAD(P) are considered to be vestiges of early evolution that proceeded via the RNA/peptide world.⁸⁹ This is a plausible example of how an ancestral protein started to oligomerize and exhibit a catalytic activity.

Consistent with this hypothesis, the DfrB mechanism is not entropically and enthalpically optimized. Typically, substrate binding to an enzyme is accompanied by release of water and is mediated by specific ligand-enzyme interactions, thereby increasing entropy and providing favorable enthalpy.⁹⁰ This is the case for NADPH binding, where water molecules are released to favor direct contacts with DfrB1.⁹¹ However, water uptake is observed upon DHF binding to DfrB1, due to water molecules mediating the new interactions; this is a hallmark of a poorly evolved mechanism.⁹¹ Furthermore, whereas NADPH remains rigid in the ternary complex, the DHF glutamate tail fluctuates between the symmetrical Lys32 at the tunnel mouth (Figure 2.3).⁴⁹ This motion of the glutamate tail results in puckering of the pterin ring that promotes the formation of the transition state.⁴⁹ These observations are consistent with the DfrB not having specifically adapted to bind DHF in particular, yet exploiting the atypical active-site symmetry to promote its reduction by the unusual means of substrate disorder-assisted catalysis.⁶⁴

2.9 When catalysis means survival: the emergence of a powerful antibiotic resistance mechanism

The synthetic antimicrobial trimethoprim (TMP) was first introduced into clinical practice in England, in 1962.^{17,19,92} Because of its efficacy and since it is well tolerated, TMP rapidly became an antimicrobial of choice to treat common urinary tract pathogens, among other infections. In 1972, the first member of the DfrB family was identified in TMP-resistant pathogenic bacteria isolated from a clinical urine sample (Figure 2.6).¹⁸ Similarly to DfrB1, DfrB2 was also discovered in a TMP-resistant clinical sample, and the identification of DfrB genes in clinical isolates became more common.⁹³⁻⁹⁷ TMP specifically targets Folate,

the chromosomal bacterial dihydrofolate reductase but is ineffective against the evolutionarily unrelated DfrB.¹⁴

TMP is in wide clinical use today, and is broadly used in veterinary settings for treatment and prevention of bacterial infections in livestock and aquaculture.^{98–100} Advances in genomics and bioinformatics have recently led to the observation that DfrB genes co-occur with other antimicrobial resistance genes and flanked by markers of genetic mobility.^{101–103} These reports unequivocally demonstrate that DfrB genes are being disseminated among pathogenic microbes in contexts that promote multi-drug resistance.

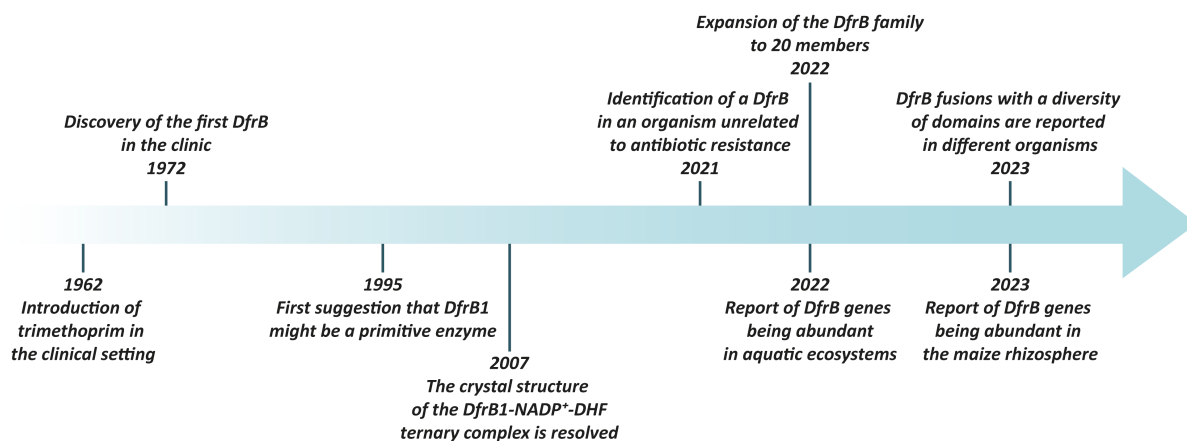


Figure 2.6. Timeline of the key discoveries concerning the DfrB enzymes.

DfrB1 through DfrB7, the first members of the DfrB family to be reported, were identified in such multi-drug resistant contexts, in samples associated with human activities such as clinical, veterinary or wastewater samples. Finding DfrB in TMP-resistant contexts has led to an additional important question: do they also occur in environmental contexts that are unrelated to human activity? The first indication that DfrB genes exist in contexts unrelated to human activities was our discovery of DfrB11 in a *β-proteobacteria* from deep terrestrial subsurface sediments in Japan.^{103,104} DfrB11 shares 87% DNA identity with the closest DfrB member known at the time, can procure TMP resistance and has all the known important sequence features for activity, yet it is not found in the vicinity of other antibiotic resistance genes.

More recently, we reported an additional set of ten DfrB genes identified in *proteobacteria* from environments not directly associated with antibiotic use, such as different types of soil, growing the number of confirmed DfrB from eight (prior to 2019) to 20 (in 2022).⁴⁴ In most cases, the genes identified alongside the DfrB are unrelated to antibiotic resistance. An independent study found that DfrB genes are abundant

in the maize rhizosphere (microbes associated with the roots of corn), suggesting that DfrB are intrinsic to this environment.¹⁰⁵ Natural freshwater environments were also shown to be a major reservoir of DfrB variants.^{106,107} In all these cases, DfrB were identified in environments that have presumably not been exposed to TMP such that selective pressure does not justify their presence. We therefore hypothesized that the DfrB in these environmental organisms were selected for a function other than to provide TMP resistance; whether they procure an evolutionary advantage by serving as an additional dihydrofolate reductase or have another, yet undiscovered function remains to be determined.⁴⁴

A factor contributing to the fact that their native function is unknown is that the DfrB are evolutionarily unrelated to any other characterized protein. Interestingly, we have identified proteins of unknown function from bacteriophages and *proteobacteria* that contain a DfrB domain – with all sequence features necessary to yield an active enzyme – within larger proteins of length ranging from 167 to 463 residues.³⁶ We have shown that these putative proteins, which have ~56% local sequence identity to DfrB1, can catalyze the reduction of dihydrofolate in a manner similar to DfrB. We also demonstrated that the two classes of DfrB inhibitors we designed can inhibit these homologues, confirming the similarity of their active-site pore. However, the native function of the DfrB domain is unknown in these larger proteins.

2.10 Conclusions and perspectives

The line between binding and catalysis can be blurry. As demonstrated above, the DfrB enzyme system illustrates strategies that are exploited by a simple 56-amino acid protein fold to catalyze imine reduction in a selective manner. The strategies that tipped the balance from a typical SH3-type binding fold to a catalyst are (1) tetramerization to increase the binding surface area while imposing constraints, (2) proximity-based catalysis that provides positive cooperativity for backbone-assisted binding of DHF to the DfrB:NADPH complex and (3) substrate disorder-assisted catalysis whereby the active-site symmetry is used in an advantageous manner to promote the formation of the transition state. As a result, and despite their modest catalytic efficiency and their unusual symmetry that limits further evolutionary optimization, DfrB enzymes catalyze the reduction of DHF to THF with an efficiency that has resulted in their mobilization into pathogenic bacteria to procure resistance to trimethoprim.

The suboptimal design of the four-fold symmetrical active site coupled with the lack of optimization of the catalytic mechanism underscore the ‘primitive’ nature of DfrB enzymes, likely derived from an ancestral binding module. This highlights the dynamic boundary between binding and catalysis in enzyme evolution.

We are currently exploring the wider sequence space occupied by the DfrB domain. We anticipate that acquiring greater knowledge of the relationships between the DfrB enzymes and their evolutionary homologues will shed light on the essential features that make this SH3 domain functional as a TMP-

resistant catalyst for the synthesis of THF, to provide a glimpse of how nature stumbles on proteins that provide unexpected value.

2.11 Conflicts of interest

There are not conflicts to declare.

2.12 Acknowledgements

The authors gratefully acknowledge Alma Carolina Sanchez-Rocha and Stella Cellier-Goetghebeur for insightful discussions and proofreading. This work was supported by the Natural Science and Engineering Research Council of Canada (NSERC) discovery grant RGPIN-N-2018-04686 and the Canada Research Chair in Engineering of Applied Proteins (J.N.P.). C.L.-S.-D was supported by scholarships from NSERC and Hydro-Québec.

2.13 References

- (1) Horvat, C. M.; Wolfenden, R. V. A Persistent Pesticide Residue and the Unusual Catalytic Proficiency of a Dehalogenating Enzyme. *Proc. Natl. Acad. Sci.* **2005**, *102* (45), 16199–16202. <https://doi.org/10.1073/pnas.0508176102>.
- (2) Samson, R.; Deutch, J. M. Diffusion-Controlled Reaction Rate to a Buried Active Site. *J. Chem. Phys.* **1978**, *68* (1), 285. <https://doi.org/10.1063/1.435494>.
- (3) Schurr, J. M.; Schmitz, K. S. Orientation Constraints and Rotational Diffusion in Bimolecular Solution Kinetics. A Simplification. *J. Phys. Chem.* **1976**, *80* (17), 1934–1936. <https://doi.org/10.1021/j100558a026>.
- (4) Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davidi, D.; Tawfik, D. S.; Milo, R. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **2011**, *50* (21), 4402–4410. <https://doi.org/10.1021/bi2002289>.
- (5) Snider, M. G.; Temple, B. S.; Wolfenden, R. The Path to the Transition State in Enzyme Reactions: A Survey of Catalytic Efficiencies. *J. Phys. Org. Chem.* **2004**, *17* (6–7), 586–591. <https://doi.org/10.1002/poc.761>.
- (6) Bugg, T. D. H. Enzyme Catalysis: Chemical Strategies. In *Wiley Encyclopedia of Chemical Biology*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2008; p wecb154. <https://doi.org/10.1002/9780470048672.wecb154>.
- (7) Tan, K. L. Temporary Intramolecularity. *Nat. Chem.* **2012**, *4* (4), 253–254. <https://doi.org/10.1038/nchem.1308>.
- (8) Krahn, J. M.; Jackson, M. R.; DeRose, E. F.; Howell, E. E.; London, R. E. Crystal Structure of a Type II Dihydrofolate Reductase Catalytic Ternary Complex [†]. *Biochemistry* **2007**, *46* (51), 14878–14888. <https://doi.org/10.1021/bi701532r>.
- (9) Alonso, H.; Gillies, M. B.; Cummins, P. L.; Bliznyuk, A. A.; Gready, J. E. Multiple Ligand-Binding Modes in Bacterial R67 Dihydrofolate Reductase. *J. Comput. Aided Mol. Des.* **2005**, *19* (3), 165–

187. <https://doi.org/10.1007/s10822-005-3693-6>.
- (10) Park, H.; Zhuang, P.; Nichols, R.; Howell, E. E. Mechanistic Studies of R67 Dihydrofolate Reductase. *J. Biol. Chem.* **1997**, *272* (4), 2252–2258. <https://doi.org/10.1074/jbc.272.4.2252>.
 - (11) Mhashal, A. R.; Pshetitsky, Y.; Cheatum, C. M.; Kohen, A.; Major, D. T. Evolutionary Effects on Bound Substrate pK_a in Dihydrofolate Reductase. *J. Am. Chem. Soc.* **2018**, *140* (48), 16650–16660. <https://doi.org/10.1021/jacs.8b09089>.
 - (12) Deng, H.; Callender, R.; Howell, E. Vibrational Structure of Dihydrofolate Bound to R67 Dihydrofolate Reductase. *J. Biol. Chem.* **2001**, *276* (52), 48956–48960. <https://doi.org/10.1074/jbc.M105107200>.
 - (13) Mhashal, A. R.; Major, D. T. Temperature-Dependent Kinetic Isotope Effects in R67 Dihydrofolate Reductase from Path-Integral Simulations. *J. Phys. Chem. B* **2021**, *125* (5), 1369–1377. <https://doi.org/10.1021/acs.jpcc.0c10318>.
 - (14) Howell, E. E. Searching Sequence Space: Two Different Approaches to Dihydrofolate Reductase Catalysis. *ChemBioChem* **2005**, *6* (4), 590–600. <https://doi.org/10.1002/cbic.200400237>.
 - (15) Matthews, D. A.; Smith, S. L.; Bacanari, D. P.; Burchall, J. J.; Oatley, S. J.; Kraut, J. Crystal Structure of a Novel Trimethoprim-Resistant Dihydrofolate Reductase Specified in Escherichia Coli by R-Plasmid R67. *Biochemistry* **1986**, *25* (15), 4194–4204. <https://doi.org/10.1021/bi00363a005>.
 - (16) Alonso, H.; Gready, J. E. Integron-Sequestered Dihydrofolate Reductase: A Recently Redeployed Enzyme. *Trends Microbiol.* **2006**, *14* (5), 236–242. <https://doi.org/10.1016/j.tim.2006.03.003>.
 - (17) Pattishall, K. H.; Acar, J.; Burchall, J. J.; Goldstein, F. W.; Harvey, R. J. Two Distinct Types of Trimethoprim-Resistant Dihydrofolate Reductase Specified by R-Plasmids of Different Compatibility Groups. *J. Biol. Chem.* **1977**, *252* (7), 2319–2323.
 - (18) Fleming, M. P.; Datta, N.; Gruneberg, R. N. Trimethoprim Resistance Determined by R Factors. *BMJ* **1972**, *1* (5802), 726–728. <https://doi.org/10.1136/bmj.1.5802.726>.
 - (19) Amyes, S. G. B.; Smith, J. T. R-Factor Trimethoprim Resistance Mechanism: An Insusceptible Target Site. *Biochem. Biophys. Res. Commun.* **1974**, *58* (2), 412–418. [https://doi.org/10.1016/0006-291X\(74\)90380-5](https://doi.org/10.1016/0006-291X(74)90380-5).
 - (20) Martinez, M. A.; Pezo, V.; Marlière, P.; Wain-Hobson, S. Exploring the Functional Robustness of an Enzyme by in Vitro Evolution. *EMBO J.* **1996**, *15* (6), 1203–1210.
 - (21) Mager, P. P. The Active Site of HIV-1 Protease. *Med. Res. Rev.* **2001**, *21* (4), 348–353. <https://doi.org/10.1002/med.1012>.
 - (22) Bhaumik, P.; Schmitz, W.; Hassinen, A.; Hiltunen, J. K.; Conzelmann, E.; Wierenga, R. K. The Catalysis of the 1,1-Proton Transfer by α -Methyl-Acyl-CoA Racemase Is Coupled to a Movement of the Fatty Acyl Moiety Over a Hydrophobic, Methionine-Rich Surface. *J. Mol. Biol.* **2007**, *367* (4), 1145–1161. <https://doi.org/10.1016/j.jmb.2007.01.062>.
 - (23) Narayana, N.; Matthews, D. A.; Howell, E. E.; Xuong, N. A Plasmid-Encoded Dihydrofolate

- Reductase from Trimethoprim-Resistant Bacteria Has a Novel D2-Symmetric Active Site. *Nat. Struct. Mol. Biol.* **1995**, *2* (11), 1018–1025. <https://doi.org/10.1038/nsb1195-1018>.
- (24) Kaneko, T.; Li, L.; Li, S. S.-C. The SH3 Domain- a Family of Versatile Peptide- and Protein-Recognition Module. *Front. Biosci.* **2008**, *Volume* (13), 4938. <https://doi.org/10.2741/3053>.
- (25) Mayer, B. J. The Discovery of Modular Binding Domains: Building Blocks of Cell Signalling. *Nat. Rev. Mol. Cell Biol.* **2015**, *16* (11), 691–698. <https://doi.org/10.1038/nrm4068>.
- (26) Alvarez-Carreño, C.; Penev, P. I.; Petrov, A. S.; Williams, L. D. Fold Evolution before LUCA: Common Ancestry of SH3 Domains and OB Domains. *Mol. Biol. Evol.* **2021**, *38* (11), 5134–5143. <https://doi.org/10.1093/molbev/msab240>.
- (27) Dionne, U.; Percival, L. J.; Chartier, F. J. M.; Landry, C. R.; Bisson, N. SRC Homology 3 Domains: Multifaceted Binding Modules. *Trends Biochem. Sci.* **2022**, *47* (9), 772–784. <https://doi.org/10.1016/j.tibs.2022.04.005>.
- (28) Kishan, K.; Agrawal, V. SH3-like Fold Proteins Are Structurally Conserved and Functionally Divergent. *Curr. Protein Pept. Sci.* **2005**, *6* (2), 143–150. <https://doi.org/10.2174/1389203053545444>.
- (29) Bodenreider, C.; Kellershohn, N.; Goldberg, M. E.; Méjean, A. Kinetic Analysis of R67 Dihydrofolate Reductase Folding: From the Unfolded Monomer to the Native Tetramer. *Biochemistry* **2002**, *41* (50), 14988–14999. <https://doi.org/10.1021/bi020453b>.
- (30) Méjean, A.; Bodenreider, C.; Schuerer, K.; Goldberg, M. E. Kinetic Characterization of the pH-Dependent Oligomerization of R67 Dihydrofolate Reductase. *Biochemistry* **2001**, *40* (27), 8169–8179. <https://doi.org/10.1021/bi010611j>.
- (31) Reece, L. J.; Nichols, R.; Ogden, R. C.; Howell, E. E. Construction of a Synthetic Gene for an R-Plasmid-Encoded Dihydrofolate Reductase and Studies on the Role of the N-Terminus in the Protein. *Biochemistry* **1991**, *30* (45), 10895–10904. <https://doi.org/10.1021/bi00109a013>.
- (32) West, F. W.; Seo, H.-S.; Bradrick, T. D.; Howell, E. E. Effects of Single-Tryptophan Mutations on R67 Dihydrofolate Reductase †. *Biochemistry* **2000**, *39* (13), 3678–3689. <https://doi.org/10.1021/bi992195x>.
- (33) Nichols, R.; Weaver, C. D.; Eisenstein, E.; Blakley, R. L.; Appleman, J.; Huang, T. H.; Huang, F. Y.; Howell, E. E. Titration of Histidine 62 in R67 Dihydrofolate Reductase Is Linked to a Tetramer .Tautm. Two-Dimer Equilibrium. *Biochemistry* **1993**, *32* (7), 1695–1706. <https://doi.org/10.1021/bi00058a002>.
- (34) Dam, J. Effect of Multiple Symmetries on the Association of R67 DHFR Subunits Bearing Interfacial Complementing Mutations. *Protein Sci.* **2004**, *13* (1), 1–14. <https://doi.org/10.1110/ps.03309504>.
- (35) *The Merck Index: An Encyclopedia of Chemicals, Drugs, and Biologicals*, 15th ed.; O’Neil, M. J., Heckelman, P. E., Dobbelaar, P. H., Roman, K. J., Kenny, C. M., Karaffa, L. S., Royal Society of Chemistry (Great Britain), Eds.; Royal Society of Chemistry: Cambridge, UK, 2013.
- (36) Lemay-St-Denis, C.; Alejaldre, L.; Jemouai, Z.; Lafontaine, K.; St-Aubin, M.; Hitache, K.;

- Valikhani, D.; Weerasinghe, N. W.; Létourneau, M.; Thibodeaux, C. J.; Doucet, N.; Baron, C.; Copp, J. N.; Pelletier, J. N. A Conserved SH3-like Fold in Diverse Putative Proteins Tetramerizes into an Oxidoreductase Providing an Antimicrobial Resistance Phenotype. *Philos. Trans. R. Soc. B Biol. Sci.* **2023**, *378* (1871), 20220040. <https://doi.org/10.1098/rstb.2022.0040>.
- (37) Hicks, S. N.; Smiley, R. D.; Hamilton, J. B.; Howell, E. E. Role of Ionic Interactions in Ligand Binding and Catalysis of R67 Dihydrofolate Reductase[†]. *Biochemistry* **2003**, *42* (36), 10569–10578. <https://doi.org/10.1021/bi034643d>.
- (38) Zhuang, P.; Eisenstein, E.; Howell, E. E. Equilibrium Folding Studies of Tetrameric R67 Dihydrofolate Reductase. *Biochemistry* **1994**, *33* (14), 4237–4244. <https://doi.org/10.1021/bi00180a018>.
- (39) Pitcher, W. H.; DeRose, E. F.; Mueller, G. A.; Howell, E. E.; London, R. E. NMR Studies of the Interaction of a Type II Dihydrofolate Reductase with Pyridine Nucleotides Reveal Unexpected Phosphatase and Reductase Activity. *Biochemistry* **2003**, *42* (38), 11150–11160. <https://doi.org/10.1021/bi0349874>.
- (40) Bhojane, P. P.; Duff, M. R.; Bafna, K.; Agarwal, P.; Stanley, C.; Howell, E. E. Small Angle Neutron Scattering Studies of R67 Dihydrofolate Reductase, a Tetrameric Protein with Intrinsically Disordered N-Termini. *Biochemistry* **2017**, *56* (44), 5886–5899. <https://doi.org/10.1021/acs.biochem.7b00822>.
- (41) Ebert, M. C. C. J. C.; Morley, K. L.; Volpato, J. P.; Schmitzer, A. R.; Pelletier, J. N. Asymmetric Mutations in the Tetrameric R67 Dihydrofolate Reductase Reveal High Tolerance to Active-Site Substitutions: Asymmetric Mutations in R67 Dihydrofolate Reductase. *Protein Sci.* **2015**, *24* (4), 495–507. <https://doi.org/10.1002/pro.2602>.
- (42) Toulouse, J. L.; Shi, G.; Lemay-St-Denis, C.; Ebert, M. C. C. J. C.; Deon, D.; Gagnon, M.; Ruediger, E.; Saint-Jacques, K.; Forge, D.; Vanden Eynde, J. J.; Marinier, A.; Ji, X.; Pelletier, J. N. Dual-Target Inhibitors of the Folate Pathway Inhibit Intrinsically Trimethoprim-Resistant DfrB Dihydrofolate Reductases. *ACS Med. Chem. Lett.* **2020**, *11* (11), 2261–2267. <https://doi.org/10.1021/acsmchemlett.0c00393>.
- (43) Zhuang, P.; Yin, M.; Holland, J. C.; Peterson, C. B.; Howell, E. E. Artificial Duplication of the R67 Dihydrofolate Reductase Gene to Create Protein Asymmetry. Effects on Protein Activity and Folding. *J. Biol. Chem.* **1993**, *268* (30), 22672–22679. [https://doi.org/10.1016/S0021-9258\(18\)41580-3](https://doi.org/10.1016/S0021-9258(18)41580-3).
- (44) Cellier-Goetghebeur, S.; Lafontaine, K.; Lemay-St-Denis, C.; Tsamo, P.; Bonneau-Burke, A.; Copp, J. N.; Pelletier, J. N. Discovery of Highly Trimethoprim-Resistant DfrB Dihydrofolate Reductases in Diverse Environmental Settings Suggests an Evolutionary Advantage Unrelated to Antibiotic Resistance. *Antibiotics* **2022**, *11* (12), 1768. <https://doi.org/10.3390/antibiotics11121768>.
- (45) Cisneros, A. F.; Gagnon-Arsenault, I.; Dubé, A. K.; Després, P. C.; Kumar, P.; Lafontaine, K.; Pelletier, J. N.; Landry, C. R. Epistasis between Promoter Activity and Coding Mutations Shapes Gene Evolvability. *Sci. Adv.* **2023**, *9* (5), eadd9109. <https://doi.org/10.1126/sciadv.add9109>.
- (46) Fuente-Gómez, G. J.; Kellum, C. L.; Miranda, A. C.; Duff, M. R.; Howell, E. E. Differentiation of

- the Binding of Two Ligands to a Tetrameric Protein with a Single Symmetric Active Site by ¹⁹F NMR. *Protein Sci.* **2021**, *30* (2), 477–484. <https://doi.org/10.1002/pro.4007>.
- (47) Nam, H.; Lewis, N. E.; Lerman, J. A.; Lee, D.-H.; Chang, R. L.; Kim, D.; Palsson, B. O. Network Context and Selection in the Evolution to Enzyme Specificity. *Science* **2012**, *337* (6098), 1101–1104. <https://doi.org/10.1126/science.1216861>.
- (48) Bystroff, C.; Oatley, S. J.; Kraut, J. Crystal Structures of Escherichia Coli Dihydrofolate Reductase: The NADP⁺ Holoenzyme and the Folate .Cntdot. NADP⁺ Ternary Complex. Substrate Binding and a Model for the Transition State. *Biochemistry* **1990**, *29* (13), 3263–3277. <https://doi.org/10.1021/bi00465a018>.
- (49) Kamath, G.; Howell, E. E.; Agarwal, P. K. The Tail Wagging the Dog: Insights into Catalysis in R67 Dihydrofolate Reductase. *Biochemistry* **2010**, *49* (42), 9078–9088. <https://doi.org/10.1021/bi1007222>.
- (50) Jackson, M.; Chopra, S.; Smiley, R. D.; Maynard, P. O.; Rosowsky, A.; London, R. E.; Levy, L.; Kalman, T. I.; Howell, E. E. Calorimetric Studies of Ligand Binding in R67 Dihydrofolate Reductase. *Biochemistry* **2005**, *44* (37), 12420–12433. <https://doi.org/10.1021/bi050881s>.
- (51) Chopra, S.; Lynch, R.; Kim, S.-H.; Jackson, M.; Howell, E. E. Effects of Temperature and Viscosity on R67 Dihydrofolate Reductase Catalysis. *Biochemistry* **2006**, *45* (21), 6596–6605. <https://doi.org/10.1021/bi052504l>.
- (52) Narayana, N. High-Resolution Structure of a Plasmid-Encoded Dihydrofolate Reductase: Pentagonal Network of Water Molecules in the *D*₂-Symmetric Active Site. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62* (7), 695–706. <https://doi.org/10.1107/S0907444906014764>.
- (53) Bradrick, T. D.; Beechem, J. M.; Howell, E. E. Unusual Binding Stoichiometries and Cooperativity Are Observed during Binary and Ternary Complex Formation in the Single Active Pore of R67 Dihydrofolate Reductase, a *D*₂ Symmetric Protein. *Biochemistry* **1996**, *35* (35), 11414–11424. <https://doi.org/10.1021/bi960205d>.
- (54) Strader, M. B.; Smiley, R. D.; Stinnett, L. G.; VerBerkmoes, N. C.; Howell, E. E. Role of S65, Q67, I68, and Y69 Residues in Homotetrameric R67 Dihydrofolate Reductase †. *Biochemistry* **2001**, *40* (38), 11344–11352. <https://doi.org/10.1021/bi0110544>.
- (55) Smiley, R. D.; Stinnett, L. G.; Saxton, A. M.; Howell, E. E. Breaking Symmetry: Mutations Engineered into R67 Dihydrofolate Reductase, a *D*₂ Symmetric Homotetramer Possessing a Single Active Site Pore. *Biochemistry* **2002**, *41* (52), 15664–15675. <https://doi.org/10.1021/bi026676j>.
- (56) Park, H.; Bradrick, T. D.; Howell, E. E. A Glutamine 67--> Histidine Mutation in Homotetrameric R67 Dihydrofolate Reductase Results in Four Mutations per Single Active Site Pore and Causes Substantial Substrate and Cofactor Inhibition. *Protein Eng. Des. Sel.* **1997**, *10* (12), 1415–1424. <https://doi.org/10.1093/protein/10.12.1415>.
- (57) Dion, A.; Linn, C. E.; Bradrick, T. D.; Georghiou, S.; Howell, E. E. How Do Mutations at Phenylalanine-153 and Isoleucine-155 Partially Suppress the Effects of the Aspartate-27 -> Serine Mutation in Escherichia Coli Dihydrofolate Reductase? *Biochemistry* **1993**, *32* (13), 3479–3487.

<https://doi.org/10.1021/bi00064a036>.

- (58) Wan, Q.; Bennett, B. C.; Wilson, M. A.; Kovalevsky, A.; Langan, P.; Howell, E. E.; Dealwis, C. Toward Resolving the Catalytic Mechanism of Dihydrofolate Reductase Using Neutron and Ultrahigh-Resolution X-Ray Crystallography. *Proc. Natl. Acad. Sci.* **2014**, *111* (51), 18225–18230. <https://doi.org/10.1073/pnas.1415856111>.
- (59) Liu, C. T.; Francis, K.; Layfield, J. P.; Huang, X.; Hammes-Schiffer, S.; Kohen, A.; Benkovic, S. J. *Escherichia Coli* Dihydrofolate Reductase Catalyzed Proton and Hydride Transfers: Temporal Order and the Roles of Asp27 and Tyr100. *Proc. Natl. Acad. Sci.* **2014**, *111* (51), 18231–18236. <https://doi.org/10.1073/pnas.1415940111>.
- (60) Howell, E. E.; Villafranca, J. E.; Warren, M. S.; Oatley, S. J.; Kraut, J. Functional Role of Aspartic Acid-27 in Dihydrofolate Reductase Revealed by Mutagenesis. *Science* **1986**, *231* (4742), 1123–1128. <https://doi.org/10.1126/science.3511529>.
- (61) Wan, Q.; Bennett, B. C.; Wymore, T.; Li, Z.; Wilson, M. A.; Brooks, C. L.; Langan, P.; Kovalevsky, A.; Dealwis, C. G. Capturing the Catalytic Proton of Dihydrofolate Reductase: Implications for General Acid–Base Catalysis. *ACS Catal.* **2021**, *11* (9), 5873–5884. <https://doi.org/10.1021/acscatal.1c00417>.
- (62) Fierke, C. A.; Johnson, K. A.; Benkovic, S. J. Construction and Evaluation of the Kinetic Scheme Associated with Dihydrofolate Reductase from *Escherichia Coli*. *Biochemistry* **1987**, *26* (13), 4085–4092. <https://doi.org/10.1021/bi00387a052>.
- (63) Schnell, J. R.; Dyson, H. J.; Wright, P. E. Structure, Dynamics, and Catalytic Function of Dihydrofolate Reductase. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33* (1), 119–140. <https://doi.org/10.1146/annurev.biophys.33.110502.133613>.
- (64) Duff, M. R.; Chopra, S.; Strader, M. B.; Agarwal, P. K.; Howell, E. E. Tales of Dihydrofolate Binding to R67 Dihydrofolate Reductase. *Biochemistry* **2016**, *55* (1), 133–145. <https://doi.org/10.1021/acs.biochem.5b00981>.
- (65) Hicks, S. N.; Smiley, R. D.; Stinnett, L. G.; Minor, K. H.; Howell, E. E. Role of Lys-32 Residues in R67 Dihydrofolate Reductase Probed by Asymmetric Mutations. *J. Biol. Chem.* **2004**, *279* (45), 46995–47002. <https://doi.org/10.1074/jbc.M404484200>.
- (66) Charlton, P. A.; Young, D. W.; Birdsall, B.; Feeney, J.; Roberts, G. C. K. Stereochemistry of Reduction of Folic Acid Using Dihydrofolate Reductase. *J. Chem. Soc. Chem. Commun.* **1979**, No. 20, 922. <https://doi.org/10.1039/c39790000922>.
- (67) Castillo, R.; Andrés, J.; Moliner, V. Catalytic Mechanism of Dihydrofolate Reductase Enzyme. A Combined Quantum-Mechanical/Molecular-Mechanical Characterization of Transition State Structure for the Hydride Transfer Step. *J. Am. Chem. Soc.* **1999**, *121* (51), 12140–12147. <https://doi.org/10.1021/ja9843019>.
- (68) Brito, R. M. M.; Reddick, R.; Bennett, G. N.; Rudolph, F. B.; Rosevear, P. R. Characterization and Stereochemistry of Cofactor Oxidation by a Type II Dihydrofolate Reductase. *Biochemistry* **1990**, *29* (42), 9825–9831. <https://doi.org/10.1021/bi00494a011>.

- (69) Schmitzer, A. R.; Lépine, F.; Pelletier, J. N. Combinatorial Exploration of the Catalytic Site of a Drug-Resistant Dihydrofolate Reductase: Creating Alternative Functional Configurations. *Protein Eng. Des. Sel.* **2004**, *17* (11), 809–819. <https://doi.org/10.1093/protein/gzh090>.
- (70) Smith, S. L.; Burchall, J. J. Alpha-Pyridine Nucleotides as Substrates for a Plasmid-Specified Dihydrofolate Reductase. *Proc. Natl. Acad. Sci.* **1983**, *80* (15), 4619–4623. <https://doi.org/10.1073/pnas.80.15.4619>.
- (71) Bastien, D.; Ebert, M. C. C. J. C.; Forge, D.; Toulouse, J.; Kadnikova, N.; Perron, F.; Mayence, A.; Huang, T. L.; Vanden Eynde, J. J.; Pelletier, J. N. Fragment-Based Design of Symmetrical Bis-Benzimidazoles as Selective Inhibitors of the Trimethoprim-Resistant, Type II R67 Dihydrofolate Reductase. *J. Med. Chem.* **2012**, *55* (7), 3182–3192. <https://doi.org/10.1021/jm201645r>.
- (72) Toulouse, J. L.; Yachnin, B. J.; Ruediger, E. H.; Deon, D.; Gagnon, M.; Saint-Jacques, K.; Ebert, M. C. C. J. C.; Forge, D.; Bastien, D.; Colin, D. Y.; Vanden Eynde, J. J.; Marinier, A.; Berghuis, A. M.; Pelletier, J. N. Structure-Based Design of Dimeric Bisbenzimidazole Inhibitors to an Emergent Trimethoprim-Resistant Type II Dihydrofolate Reductase Guides the Design of Monomeric Analogues. *ACS Omega* **2019**, *4* (6), 10056–10069. <https://doi.org/10.1021/acsomega.9b00640>.
- (73) Jensen, R. A. Enzyme Recruitment in Evolution of New Function. *Annu. Rev. Microbiol.* **1976**, *30* (1), 409–425. <https://doi.org/10.1146/annurev.mi.30.100176.002205>.
- (74) Yčas, M. On Earlier States of the Biochemical System. *J. Theor. Biol.* **1974**, *44* (1), 145–160. [https://doi.org/10.1016/S0022-5193\(74\)80035-4](https://doi.org/10.1016/S0022-5193(74)80035-4).
- (75) Khersonsky, O.; Roodveldt, C.; Tawfik, D. Enzyme Promiscuity: Evolutionary and Mechanistic Aspects. *Curr. Opin. Chem. Biol.* **2006**, *10* (5), 498–508. <https://doi.org/10.1016/j.cbpa.2006.08.011>.
- (76) Longo, L. M.; Petrović, D.; Kamerlin, S. C. L.; Tawfik, D. S. Short and Simple Sequences Favored the Emergence of N-Helix Phospho-Ligand Binding Sites in the First Enzymes. *Proc. Natl. Acad. Sci.* **2020**, *117* (10), 5310–5318. <https://doi.org/10.1073/pnas.1911742117>.
- (77) Caetano-Anollés, G.; Yafremava, L. S.; Gee, H.; Caetano-Anollés, D.; Kim, H. S.; Mitterthal, J. E. The Origin and Evolution of Modern Metabolism. *Int. J. Biochem. Cell Biol.* **2009**, *41* (2), 285–297. <https://doi.org/10.1016/j.biocel.2008.08.022>.
- (78) Caetano-Anollés, G.; Kim, H. S.; Mitterthal, J. E. The Origin of Modern Metabolic Networks Inferred from Phylogenomic Analysis of Protein Architecture. *Proc. Natl. Acad. Sci.* **2007**, *104* (22), 9358–9363. <https://doi.org/10.1073/pnas.0701214104>.
- (79) David, L. A.; Alm, E. J. Rapid Evolutionary Innovation during an Archaeal Genetic Expansion. *Nature* **2011**, *469* (7328), 93–96. <https://doi.org/10.1038/nature09649>.
- (80) Ferla, M. P.; Brewster, J. L.; Hall, K. R.; Evans, G. B.; Patrick, W. M. Primordial-like Enzymes from Bacteria with Reduced Genomes. *Mol. Microbiol.* **2017**, *105* (4), 508–524. <https://doi.org/10.1111/mmi.13737>.
- (81) Makarov, M.; Meng, J.; Tretyachenko, V.; Srb, P.; Březinová, A.; Giacobelli, V. G.; Bednářová, L.; Vondrášek, J.; Dunker, A. K.; Hlouchová, K. Enzyme Catalysis Prior to Aromatic Residues: Reverse Engineering of a dephospho-CoA Kinase. *Protein Sci.* **2021**, *30* (5), 1022–1034.

<https://doi.org/10.1002/pro.4068>.

- (82) Solan, R.; Pereira, J.; Lupas, A. N.; Kolodny, R.; Ben-Tal, N. Gram-Negative Outer-Membrane Proteins with Multiple β -Barrel Domains. *Proc. Natl. Acad. Sci.* **2021**, *118* (31), e2104059118. <https://doi.org/10.1073/pnas.2104059118>.
- (83) Isildayancan, K.; Kessel, A.; Solan, R.; Kolodny, R.; Ben-Tal, N. *Proteins with Multiple G Protein-Coupled Receptor Domains*; preprint; Bioinformatics, 2022. <https://doi.org/10.1101/2022.07.26.501653>.
- (84) Feng, J.; Grubbs, J.; Dave, A.; Goswami, S.; Horner, C. G.; Howell, E. E. Radical Redesign of a Tandem Array of Four R67 Dihydrofolate Reductase Genes Yields a Functional, Folded Protein Possessing 45 Substitutions. *Biochemistry* **2010**, *49* (34), 7384–7392. <https://doi.org/10.1021/bi1005943>.
- (85) Bradrick, T. D.; Shattuck, C.; Strader, M. B.; Wicker, C.; Eisenstein, E.; Howell, E. E. Redesigning the Quaternary Structure of R67 Dihydrofolate Reductase. *J. Biol. Chem.* **1996**, *271* (45), 28031–28037. <https://doi.org/10.1074/jbc.271.45.28031>.
- (86) Stinnett, L. G.; Smiley, R. D.; Hicks, S. N.; Howell, E. E. “Catch 222,” the Effects of Symmetry on Ligand Binding and Catalysis in R67 Dihydrofolate Reductase as Determined by Mutations at Tyr-69. *J. Biol. Chem.* **2004**, *279* (45), 47003–47009. <https://doi.org/10.1074/jbc.M404485200>.
- (87) Jimenez-Morales, D.; Liang, J.; Eisenberg, B. Ionizable Side Chains at Catalytic Active Sites of Enzymes. *Eur. Biophys. J.* **2012**, *41* (5), 449–460. <https://doi.org/10.1007/s00249-012-0798-4>.
- (88) Narunsky, A.; Kessel, A.; Solan, R.; Alva, V.; Kolodny, R.; Ben-Tal, N. On the Evolution of Protein–Adenine Binding. *Proc. Natl. Acad. Sci.* **2020**, *117* (9), 4701–4709. <https://doi.org/10.1073/pnas.1911349117>.
- (89) White, H. B. Coenzymes as Fossils of an Earlier Metabolic State. *J. Mol. Evol.* **1976**, *7* (2), 101–104. <https://doi.org/10.1007/BF01732468>.
- (90) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist’s Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53* (14), 5061–5084. <https://doi.org/10.1021/jm100112j>.
- (91) Chopra, S.; Dooling, R. M.; Horner, C. G.; Howell, E. E. A Balancing Act between Net Uptake of Water during Dihydrofolate Binding and Net Release of Water upon NADPH Binding in R67 Dihydrofolate Reductase. *J. Biol. Chem.* **2008**, *283* (8), 4690–4698. <https://doi.org/10.1074/jbc.M709443200>.
- (92) Noall, E. W. P.; Sowards, H. F. G.; Waterworth, P. M. Successful Treatment of a Case of Proteus Septicaemia. *BMJ* **1962**, *2* (5312), 1101–1102. <https://doi.org/10.1136/bmj.2.5312.1101>.
- (93) Huovinen, P.; Sundström, L.; Swedberg, G.; Sköld, O. Trimethoprim and Sulfonamide Resistance. *Antimicrob. Agents Chemother.* **1995**, *39* (2), 279–289. <https://doi.org/10.1128/AAC.39.2.279>.
- (94) Fling, M. E.; Walton, L.; Elwell, L. P. Monitoring of Plasmid-Encoded, Trimethoprim-Resistant Dihydrofolate Reductase Genes: Detection of a New Resistant Enzyme. *Antimicrob. Agents Chemother.* **1982**, *22* (5), 882–888. <https://doi.org/10.1128/AAC.22.5.882>.

- (95) Toulouse, J. L.; Edens, T. J.; Alejaldre, L.; Manges, A. R.; Pelletier, J. N. Integron-Associated DfrB4, a Previously Uncharacterized Member of the Trimethoprim-Resistant Dihydrofolate Reductase B Family, Is a Clinically Identified Emergent Source of Antibiotic Resistance. *Antimicrob. Agents Chemother.* **2017**, *61* (5), e02665-16. <https://doi.org/10.1128/AAC.02665-16>.
- (96) Grape, M.; Sundström, L.; Kronvall, G. New Dfr2 Gene as a Single-Gene Cassette in a Class 1 Integron from a Trimethoprim-Resistant Escherichia Coli Isolate. *Microb. Drug Resist. Larchmt. N* **2003**, *9* (4), 317–322. <https://doi.org/10.1089/107662903322762734>.
- (97) Grape, M.; Farra, A.; Kronvall, G.; Sundström, L. Integrons and Gene Cassettes in Clinical Isolates of Co-Trimoxazole-Resistant Gram-Negative Bacteria. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* **2005**, *11* (3), 185–192. <https://doi.org/10.1111/j.1469-0691.2004.01059.x>.
- (98) Cuong, N.; Padungtod, P.; Thwaites, G.; Carrique-Mas, J. Antimicrobial Usage in Animal Production: A Review of the Literature with a Focus on Low- and Middle-Income Countries. *Antibiotics* **2018**, *7* (3), 75. <https://doi.org/10.3390/antibiotics7030075>.
- (99) Kadlec, K.; Kehrenberg, C.; Schwarz, S. Molecular Basis of Resistance to Trimethoprim, Chloramphenicol and Sulphonamides in Bordetella Bronchiseptica. *J. Antimicrob. Chemother.* **2005**, *56* (3), 485–490. <https://doi.org/10.1093/jac/dki262>.
- (100) Levings, R. S.; Lightfoot, D.; Elbourne, L. D. H.; Djordjevic, S. P.; Hall, R. M. New Integron-Associated Gene Cassette Encoding a Trimethoprim-Resistant DfrB-Type Dihydrofolate Reductase. *Antimicrob. Agents Chemother.* **2006**, *50* (8), 2863–2865. <https://doi.org/10.1128/AAC.00449-06>.
- (101) Karaolia, P.; Vasileiadis, S.; G. Michael, S.; G. Karpouzas, D.; Fatta-Kassinos, D. Shotgun Metagenomics Assessment of the Resistome, Mobilome, Pathogen Dynamics and Their Ecological Control Modes in Full-Scale Urban Wastewater Treatment Plants. *J. Hazard. Mater.* **2021**, *418*, 126387. <https://doi.org/10.1016/j.jhazmat.2021.126387>.
- (102) Baniga, Z.; Hounmanou, Y. M. G.; Kudirkiene, E.; Kusiluka, L. J. M.; Mdegela, R. H.; Dalsgaard, A. Genome-Based Analysis of Extended-Spectrum β -Lactamase-Producing Escherichia Coli in the Aquatic Environment and Nile Perch (*Lates Niloticus*) of Lake Victoria, Tanzania. *Front. Microbiol.* **2020**, *11*, 108. <https://doi.org/10.3389/fmicb.2020.00108>.
- (103) Lemay-St-Denis, C.; Diwan, S.-S.; Pelletier, J. N. The Bacterial Genomic Context of Highly Trimethoprim-Resistant DfrB Dihydrofolate Reductases Highlights an Emerging Threat to Public Health. *Antibiotics* **2021**, *10* (4), 433. <https://doi.org/10.3390/antibiotics10040433>.
- (104) Hernsdorf, A. W.; Amano, Y.; Miyakawa, K.; Ise, K.; Suzuki, Y.; Anantharaman, K.; Probst, A.; Burstein, D.; Thomas, B. C.; Banfield, J. F. Potential for Microbial H₂ and Metal Transformations Associated with Novel Bacteria and Archaea in Deep Terrestrial Subsurface Sediments. *ISME J.* **2017**, *11* (8), 1915–1929. <https://doi.org/10.1038/ismej.2017.39>.
- (105) Pham, D. N.; Wu, Q.; Li, M. Global Profiling of Antibiotic Resistomes in Maize Rhizospheres. *Arch. Microbiol.* **2023**, *205* (3), 89. <https://doi.org/10.1007/s00203-023-03424-z>.
- (106) Kneis, D.; Lemay-St-Denis, C.; Cellier-Goetghebeur, S.; Elena, A. X.; Berendonk, T. U.; Pelletier,

- J. N.; Heß, S. Trimethoprim Resistance in Surface and Wastewater Is Mediated by Contrasting Variants of the *dfpB* Gene. *ISME J.* **2023**, *17* (9), 1455–1466. <https://doi.org/10.1038/s41396-023-01460-7>.
- (107) Kneis, D.; Berendonk, T. U.; Forslund, S. K.; Hess, S. Antibiotic Resistance Genes in River Biofilms: A Metagenomic Approach toward the Identification of Sources and Candidate Hosts. *Environ. Sci. Technol.* **2022**, *56* (21), 14913–14922. <https://doi.org/10.1021/acs.est.2c00370>.
- (108) Okonechnikov, K.; Golosova, O.; Fursov, M. Unipro UGENE: A Unified Bioinformatics Toolkit. *Bioinformatics* **2012**, *28* (8), 1166–1167. <https://doi.org/10.1093/bioinformatics/bts091>.
- (109) Yariv, B.; Yariv, E.; Kessel, A.; Masrati, G.; Chorin, A. B.; Martz, E.; Mayrose, I.; Pupko, T.; Ben-Tal, N. Using Evolutionary Data to Make Sense of Macromolecules with a “Face-lifted” ConSurf. *Protein Sci.* **2023**, *32* (3), e4582. <https://doi.org/10.1002/pro.4582>.
- (110) Amyes, S. G. B.; Smith, J. T. The Purification and Properties of the Trimethoprim-Resistant Dihydrofolate Reductase Mediated by the R-Factor, R 388. *Eur. J. Biochem.* **1976**, *61* (2), 597–603. <https://doi.org/10.1111/j.1432-1033.1976.tb10055.x>.
- (111) Stone, S. R.; Morrison, J. F. Mechanism of Inhibition of Dihydrofolate Reductases from Bacterial and Vertebrate Sources by Various Classes of Folate Analogues. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* **1986**, *869* (3), 275–285. [https://doi.org/10.1016/0167-4838\(86\)90067-1](https://doi.org/10.1016/0167-4838(86)90067-1).
- (112) Appleman, J. R.; Howell, E. E.; Kraut, J.; Kühn, M.; Blakley, R. L. Role of Aspartate 27 in the Binding of Methotrexate to Dihydrofolate Reductase from *Escherichia Coli*. *J. Biol. Chem.* **1988**, *263* (19), 9187–9198. [https://doi.org/10.1016/S0021-9258\(19\)76524-7](https://doi.org/10.1016/S0021-9258(19)76524-7).

Chapitre 3. Exploration génomique des gènes *dfrB* cliniques

Préface

En 2019, à mon entrée au doctorat, quelques gènes *dfrB* procurant une résistance à l'antibiotique triméthoprimine aux bactéries les exprimant avaient été identifiés dans des échantillons cliniques. Toutefois, aucune vue d'ensemble sur leur distribution n'avait été offerte. La caractérisation de la distribution des gènes *dfrB* et de leur environnement génomique est essentielle pour nous permettre de comprendre dans quel(s) contexte(s) les gènes *dfrB* se trouvent présentement. L'approche entreprise ici consistait à analyser les séquences génomiques publiquement disponibles comportant un gène *dfrB* dont la capacité à procurer de la résistance au triméthoprimine a été expérimentalement confirmée. Des tendances ont pu être identifiées. Ainsi, la grande majorité des gènes *dfrB* identifiés sont intégrés dans des structures génomiques permettant la mobilité entre organismes, et on les retrouve quasi exclusivement dans des gamma-protéobactéries pathogènes. Ces observations nous ont permis d'avancer que les gènes codant pour les enzymes des DfrB représentent un risque important pour la santé publique, puisqu'on les retrouve dans une diversité de contextes génomiques de bactéries pathogènes. Ceci justifie l'intérêt de comprendre comment une telle famille d'enzymes a pu émerger et contribuer au problème mondial de la résistance aux antibiotiques.

Notamment, cette étude a fourni le premier indice suggérant que les *dfrB* peuvent se trouver dans des contextes indépendants à la résistance aux antibiotiques, ce qui sera exploré en détail dans les deux chapitres suivants.

Ce chapitre est composé d'un article publié dans le journal *Antibiotics* dans le cadre de l'édition spéciale « Antibiotic Resistance Genes: Spread and Evolution ». Des modifications mineures ont été apportées à la version incluse dans cette thèse. Cette étude a été conçue par Prof. Joelle N. Pelletier et moi-même. J'ai réalisé le travail en laboratoire, identifié et analysé les séquences génomiques et réalisé les figures. L'identification de gènes de virulence au sein des séquences ainsi que l'arbre phylogénétique ont été réalisés par Sarah-Slim Diwan, alors stagiaire au B.Sc. sous ma supervision. J'ai écrit la première version du manuscrit et Prof. Joelle N. Pelletier l'a révisé.

Article de recherche 1. The bacterial genomic context of highly trimethoprim-resistant DfrB dihydrofolate reductases highlights an emerging threat to public health

Claudèle Lemay-St-Denis^{1,2,3}, Sarah-Slim Diwan^{1,2,3} et Joelle N. Pelletier^{1,2,3,4,*}

¹ Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, QC H3T 1J4, Canada

² PROTEO, The Québec Network for Research on Protein, Function, Engineering and Applications, Québec, QC G1V 0A6, Canada

³ CGCC, Center in Green Chemistry and Catalysis, Montréal, QC H3A 0B8, Canada

⁴ Chemistry Department, Université de Montréal, Montréal, QC H2V 0B3, Canada

*Correspondence: joelle.pelletier@umontreal.ca

Antibiotics

DOI : [10.3390/antibiotics10040433](https://doi.org/10.3390/antibiotics10040433)

© MDPI 2021

3.1 Abstract

Type B dihydrofolate reductase (*dfrb*) genes were identified following the introduction of trimethoprim (TMP) in the 1960s. Although they intrinsically confer resistance to trimethoprim that is orders of magnitude greater than through other resistance mechanisms, the distribution and prevalence of these short (237 bp) genes is unknown. Indeed, this knowledge has been hampered by systematic biases in search methodologies. Here, we investigate the genomic context of *dfrbs* to gain information on their current distribution in bacterial genomes. Upon searching publicly available databases, we identify 61 sequences containing *dfrbs* genes within an analyzable genomic context. The majority (70%) of those sequences also harbor virulence genes and 97% of the *dfrbs* are found near a mobile genetic element, representing a potential risk for antibiotic resistance genes. We further identify and confirm the TMP-resistant phenotype of two new members of the family, *dfrb10* and *dfrb11*. *Dfrbs* are found both in Betaproteobacteria and Gammaproteobacteria, a majority (59%) being in *Pseudomonas aeruginosa*. Previously labelled as strictly plasmid-borne, we found 69% of *dfrbs* in the chromosome of pathogenic bacteria. Our results demonstrate that the intrinsically TMP-resistant *dfrbs* are a potential emerging threat to public health and justify closer surveillance of these genes.

3.2 Introduction

Trimethoprim (TMP) is a synthetic antimicrobial that is ranked as being highly important by the World Health Organization (WHO) ¹. TMP strongly and selectively inhibits a key enzyme in bacterial folate biosynthesis, chromosomal dihydrofolate reductase (K_i (*Escherichia coli* FoaA) = 20 pM ²), thereby effectively abolishing bacterial proliferation. This antimicrobial was initially introduced for clinical application in 1968 in combination with sulfamethoxazole, also an inhibitor of folate biosynthesis, and later used alone to treat various infections ^{3,4}. TMP is widely prescribed to adults and to children, as well as to animals, worldwide ⁵⁻⁷. In 2017, increasing concern over antibiotic resistance prompted the WHO to issue recommendations that include reduction of TMP usage with food-producing animals as well as a complete restriction of its use with animals to promote growth and for preventive measures ⁸. The goal of these recommendations is to lower the prevalence of antimicrobial resistant bacteria that could be transmitted to humans.

Multiple TMP resistance mechanisms have been reported ³. The main mechanisms are the acquisition of type A (DfrA) or type B (DfrB) TMP-resistant dihydrofolate reductases. These enzymes providing TMP resistance are expressed in addition to the TMP-sensitive, chromosomal FoaA, allowing folate synthesis and bacterial survival. The DfrA family is homologous to FoaA. It includes nearly 40 members that are 150-190 amino acids in length, similar to the FoaA family ⁹. The DfrB family currently consists of eight members; they are homotetrameric enzymes of 78 amino acids per protomer ¹⁰. Contrary to DfrAs, they are

phylogenetically and structurally unrelated to FoaA^{10,11}. Their evolutionary origin is currently unknown. DfrBs maintain full activity at clinical concentrations of TMP that fully inhibit FoaA¹². They offer TMP resistance at concentrations at least 3 orders of magnitude greater than DfrAs¹³, thus conferring TMP resistance that cannot be countered by administering TMP at higher concentrations.

DfrB1 is the first member of the DfrB family to have been reported. Over the past decades, DfrB1 has been characterized in great detail for its unique structure, its biophysical characteristics, its multimerization and its robustness¹⁴⁻¹⁸. In particular, DfrBs are distinguished from most enzymes in the fact that their single, central active site requires distinct contribution from each of the four identical protomers, creating an evolutionary conundrum^{10,17}. Recently, other members of the DfrB family have been functionally characterized by our group, showing nearly indistinguishable dihydrofolate reductase activity, high resistance to TMP and similar inhibition by recently-reported inhibitors of distinct classes¹⁹⁻²². Nonetheless, at the outset of this study, none of the widely-used databases (CARD, ARDB and ARG-ANNOT) contained all eight known *dfrb* sequences²³⁻²⁵.

The prevalence of *dfrb* in clinical and environmental samples is currently unknown. The *dfrb* genes have rarely been reported in clinical samples^{26,27}. Although this could be interpreted as the scarce presence of DfrBs in the collection of resistance genes in bacteria (resistome), it is important to note that the short 237-bp *dfrb* genes have not been routinely searched for. Until recently, gene-prediction algorithms used 300 nt as a cut-off to differentiate short non-protein-coding RNAs (ncRNA) from messenger RNAs (mRNA)²⁸. Even now, sophisticated gene prediction algorithms such as the widely used Prodigal are unlikely to identify *dfrbs* as a result of their unusual codon usage and small size, both of which are penalized²⁹.

Experimental detection of TMP-resistant DfrB enzymes has also consistently failed to detect *dfrb* genes because of the prevalence of PCR-based methods: primers specific to *dfras* are used, with few or no primers specific to the unrelated *dfrbs*^{30,31}. Fortunately, the advent of whole-genome sequencing now allows for large-scale computational screening of antimicrobial resistance gene databases without experimental biases.

According to Martínez *et al.*, the greatest public health risk is observed when antimicrobial resistance genes to widely-used antibiotics are found on mobile genetic elements (MGE) of a human pathogen³². To date, a limited number of reports have found *dfrbs* near integrases and transposases, indicative of their potential genomic mobility^{27,33}. Nonetheless, as *dfrbs* are rarely reported, their genomic context is essentially unexplored and the current public health risk that they represent is unknown.

Here, we searched publicly available databases to identify sequences containing *dfrbs*. We investigated the predicted pathogenicity of the organisms harbouring each sequence as well as the genomic context of *dfrbs*.

We found that 70% of sequences containing *dfrbs* harbor virulence genes, mostly from *Pseudomonas*, a major cause of infection in humans that is difficult to treat because of its evolved resistance³⁴. Overall, 97% of *dfrbs* are in proximity to a mobile genetic element, favoring their dissemination. Importantly, this investigation resulted in the identification of two new members of the *dfrb* family; we expressed both and confirmed their highly TMP-resistant phenotype *in vitro*. Our results demonstrate that the intrinsically TMP-resistant DfrBs can be found in a variety of contexts that are consistent with transmission of multidrug resistance and justify closer surveillance of these genes.

3.3 Results

3.3.1 Expansion of the DfrB Family

Our first objective was to determine whether additional DfrB homologues could be identified, to expand the small but rapidly growing DfrB family. Using profile hidden Markov models (HMM) of the six functionally characterized DfrBs (DfrB1-5 and DfrB7)¹⁹, we searched the TrEMBL database. In addition to confirming the presence of DfrB1-9, we also identified two genes displaying high homology to the conserved core of DfrBs yet sharing sequence identity of less than 95% to any known *dfrb* genes. They were named *dfrb10* and *dfrb11* (Table 3.1).

Table 3.1. Information and MICs on the newly identified DfrB10 and DfrB11

New name	UniprotKB accession number	Genbank accession number	Closest characterized DfrB (protein identity / DNA identity) ^a	MIC (µg/mL)
DfrB10	A0A2Z1CLP9	ALZ46148.1	DfrB3 (92% / 93%)	> 600
DfrB11	A0A2N2TNN4	PKO69073.1	DfrB3 (90% / 87%)	> 600

^a Protein sequence identity of all members of the DfrB family are reported in Table S3.1.

The *dfrB10* gene was found on the p12969-DIM mega-plasmid (0.4-Mb) from a *Pseudomonas putida* strain isolated in China in 2013 from a patient suffering from pneumonia³⁵. The *dfrB11* gene was identified in a groundwater sample at the Horonobe Underground research laboratory in Japan in 2017, in a *Betaproteobacteria* sequence³⁶. Both new DfrBs produce the same phenotype as the other DfrBs when overexpressed in *E. coli*: they confer resistance to 0.6 mg/mL TMP, the highest concentration of TMP that can be solubilized in 5% methanol (Table 3.1). This situates these genes amongst the most resistant dihydrofolate reductases known to date.

3.3.2 Identification of Bacterial Sequences

The DNA sequences of the eight previously reported *dfrbs* and the two new *dfrbs* were searched against publicly available genomic databases; we note that there is no *dfrB8* (Table S3.2)¹¹. Since our objective

was to analyze the genomic context of *dfrbs*, we retained only genomic segments that include at least 10kb both upstream and downstream from a *dfrb*. A total of 110 sequences were collected, representing 16 different bacterial species. In some cases, multiple similar sequences originated from the same BioProject; in these instances, redundant sequences were excluded from further analysis, keeping one representative sequence.

The taxonomic summary for the 61 remaining sequences is presented in Table 3.2. All sequences but one come from *Gammaproteobacteria* and include three different orders: *Aeromonadales*, *Enterobacterales* and *Pseudomonadales*. The predominant species is the clinically-relevant *Pseudomonas aeruginosa*, accounting for 36 sequences (59%). The only *Betaproteobacteria* sequence is from *Burkholderia dolosa*, isolated from a cystic fibrosis patient in the United States of America³⁷.

Table 3.2. Taxonomic classification of all strains identified that include at least one *dfrb*

Class/order/family/genus	Strain count ^a
Betaproteobacteria	1
<i>Burkholderiales</i>	1
• <i>Burkholderiaceae</i>	1
<i>Burkholderia</i>	1
Gammaproteobacteria	60 (110)
<i>Aeromonadales</i>	2
• <i>Aeromonadaceae</i>	2
<i>Aeromonas</i>	2
Enterobacterales	16 (17)
• <i>Enterobacteriaceae</i>	14 (15)
<i>Citrobacter</i>	1
<i>Enterobacter</i>	1
<i>Escherichia</i>	4
<i>Klebsiella</i>	4 (5)
<i>Salmonella</i>	4
• <i>Morganellaceae</i>	1
<i>Providencia</i>	1
• <i>Yersiniaceae</i>	1
<i>Serratia</i>	1
<i>Pseudomonadales</i>	42 (91)
• <i>Moraxellaceae</i>	1
<i>Acinetobacter</i>	1
• <i>Pseudomonadaceae</i>	41 (90)
<i>Pseudomonas</i>	41 (90)

^a Values include sequences used in the analysis after exclusion of redundancy. Values in parentheses include redundant sequences.

When available, information on the isolation source of each sample and the country of origin was compiled. Most *dfrb*-containing strains were identified in samples collected in Asia (39%), followed by Europe (20%), America (17%) and Africa (5%). The majority of strains (62%) were found within humans; the 7% found

in wastewater and 2% in hospital wastewater may also be of human origin. Surprisingly, despite intensive use of TMP in livestock, only 3% of strains from our dataset were isolated from animals. This could indicate a sampling bias from the databases. Finally, 2% of samples were identified as environmental.

Four species in the dataset (*Acinetobacter baumannii*, *Enterobacter hormaechei*, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*) are categorized as ESKAPE pathogens, accounting for 67% of the sequences. Overall, 79% of the 61 *dfrb*-containing sequences contain predicted virulence genes potentially enabling them to cause infection according to the VFDB database. Because our genomic sequence dataset included partial sequences, the fraction of sequences of pathogenic bacteria can be underestimated if virulence genes are outside of the sequenced region.

3.3.3 Analysis of the Genomic Context

We investigated the genomic context within which the *dfrbs* were found. In particular, the presence of MGEs can inform us of the capacity of the *dfrbs* to transfer to other genomes. Ever since the initial discovery of *dfrb1* (R67) and *dfrb2* (R388) on plasmids, *dfrbs* have been systematically referred to as being plasmid-borne^{38,39}. This is consistent with the importance of plasmids in acquired bacterial resistance⁴⁰. The genomic context was determined by classifying the *dfrb*-containing genomes as either plasmidic or chromosomal using the PlasFlow classification tool⁴¹. This resulted in 18 sequences being labelled “plasmids” (30%) and 43 sequences labelled “chromosomes” (70%) (Figure 3.1). The only sequence identified in *Betaproteobacteria* is chromosomal. Among the *Gammaproteobacteria*, all 16 sequences identified in *Enterobacteriales* are plasmidic, whereas the two sequences in *Aeromonadales* are chromosomal. The *Pseudomonadales* sequences are chromosomal except for two sequences: one from *P. aeruginosa* and one from *P. putida*. Only one plasmidic sequence, that from *P. rettgeri*, was labelled as being from a potential pathogen.

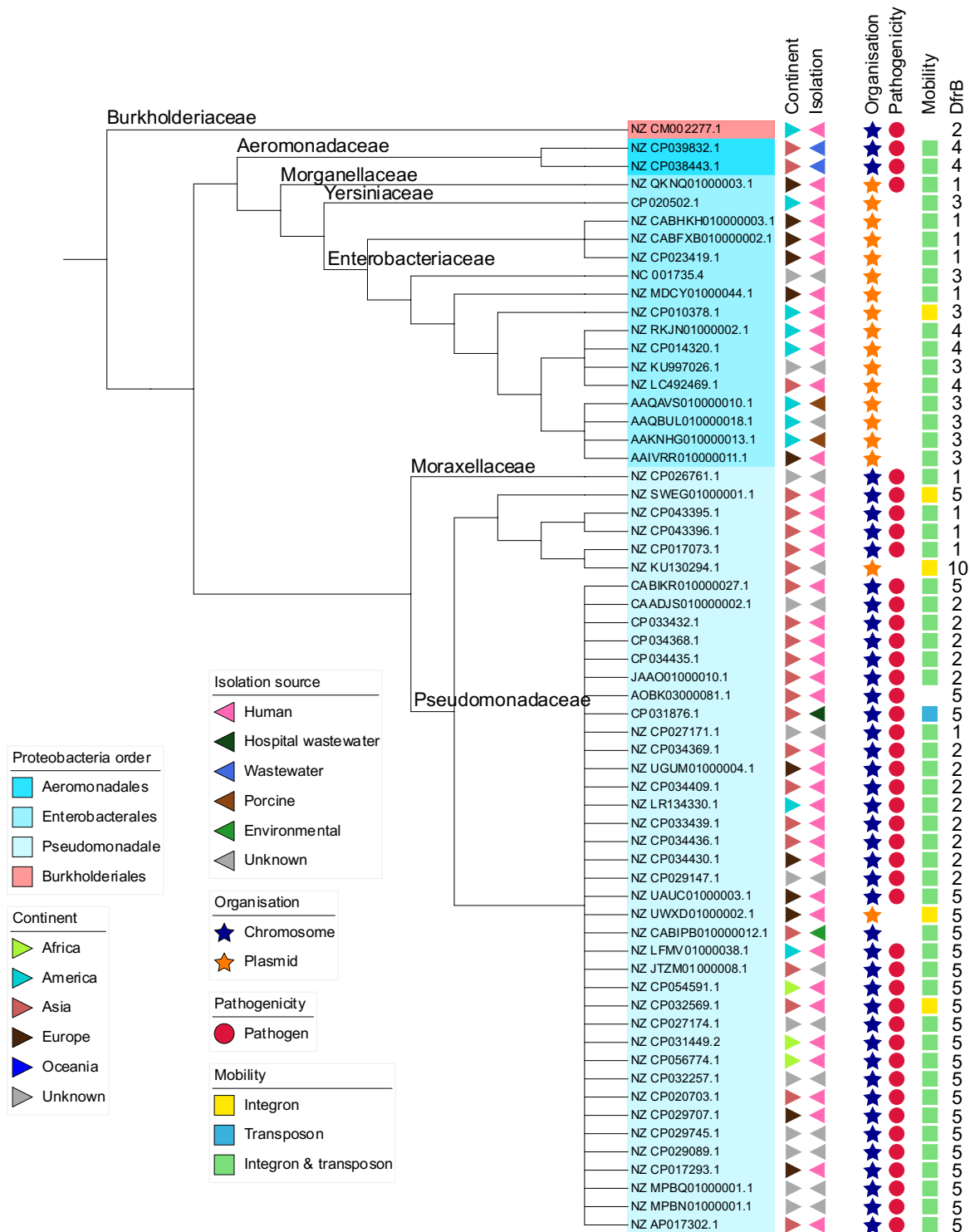


Figure 3.1. Annotated phylogenetic tree of species harboring a *dfrB*.

Taxonomic classification of order and family is followed by categorization according to GenBank information on the strain's isolation source and country of isolation. Sequences are further categorized as being located in a chromosome or a plasmid, pathogenicity of the host organism and information on mobile genetic elements. The *dfrB* gene member identified in each sequence is specified (i.e. '2' indicates *dfrB2*).

Next, we gained insight into the types of genes flanking the *dfrbs*. MGEs and other resistance genes near the *dfrbs* would define them as belonging to a multi-resistance context. A blastx search was performed, using 20kb sequence segments containing *dfrbs* as queries, against a compiled antibiotic resistance gene database (see Materials and Methods), keeping hits having at least 80% coverage and 60% identity. We first determined that the vast majority of *dfrbs* (89%) have both integrase and transposase within a 10kb window (Figure 3.1). Five sequences (NZ_CP010378.1, NZ_SWEG01000001.1, NZ_UWXD01000002.1, NZ_CP032569.1, NZ_KU130294.1) had only an integrase annotated nearby and one sequence (CP031876.1) had only a transposase annotated. Two sequences included neither; one (NZ_CM002277.1) is a chromosomal sequence from the pathogenic *Burkholderia dolosa* strain, the only *Betaproteobacteria* identified. That sequence included no additional genes related to antimicrobial resistance or genomic mobility within 10kb of its *dfrb* (Figure 3.2.a). The other sequence including no integrase or transposase (AOBK03000081.1) is a chromosomal sequence harboring virulence genes from a *P. aeruginosa* clinical isolate. It included only the rifampin-resistance gene *arr2* at a distance of 0.14kb from the *dfrb* gene.

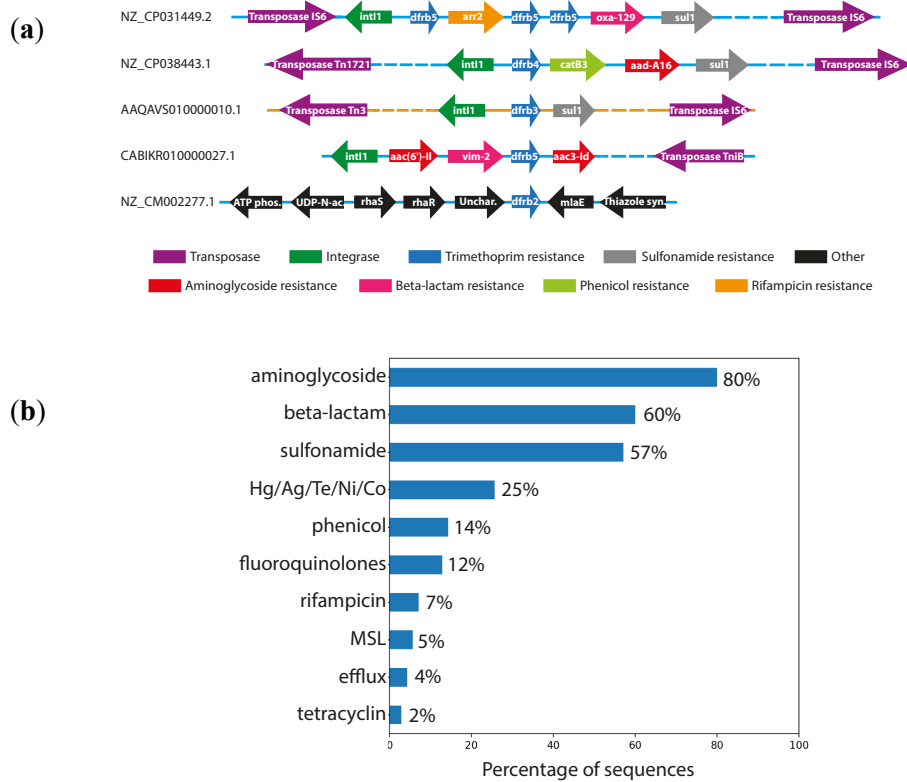


Figure 3.2.

(a) Scheme of the genomic context of five representative sequences. Blue lines represent genomic sequences, whereas the orange line represents a plasmid sequence. Dashed lines contain regions that are not represented here. **(b)** Population of *dfrb* genes accompanied by a gene conferring resistance to another antimicrobial agent, expressed as percentage. MSL: Macrolides, streptogramins, lincosamides.

The distance separating the *dfrbs* from the integrases and transposases was mapped (Figure 3.3). For *dfrb1*, *dfrb2* and *dfrb4*, we observe a large variability in those distances, suggesting diversity in their genomic context and thus diversity in the events of integration. In the case of *dfrb3* and *dfrb5*, the same gene cassette is present in a few sequences, thus the same distance between elements is mapped. For example, all eight *dfrb3* genes are found in integrons directly upstream from the class 1 integrase, marking *dfrb3* as the gene most recently integrated into the cassette. Interestingly, *dfrb3* was found only in plasmids, in *Enterobacteriales*. Six of these genes are found in the short *dfrb3/sul1* cassette. Similarly, multiple sequences held the same cassette containing *dfrb5*. There are 15 sequences with the *aac(6')-II/vim-2/dfrb5/aac(3)-Id* cassette and five with the *aac(6')-II/dfrb5/aac(3)-Id* cassette; some other cassettes were found twice.

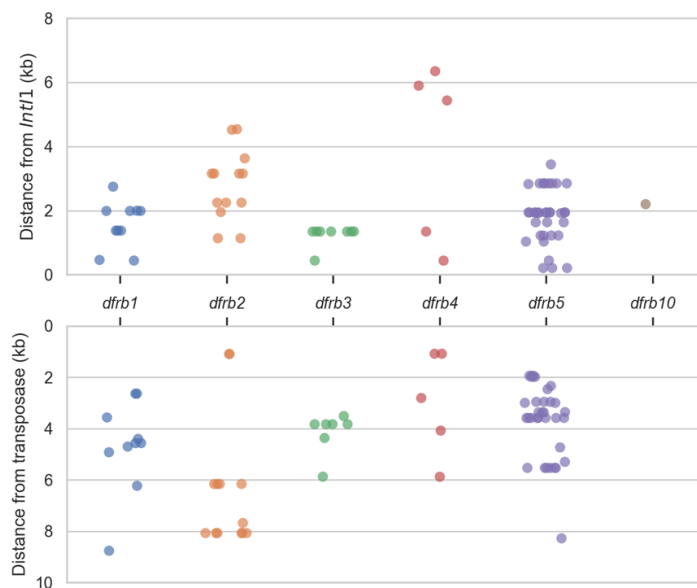


Figure 3.3. Distance between *dfrbs* and genes associated with genomic mobility.

Top panel: distance between the *dfrbs* that are downstream from a class 1 integrase. Bottom panel: distance between *dfrbs* and the closest transposase. Each dot represents one *dfrb* gene.

Finally, we mapped antimicrobial resistance genes annotated in a 10kb window on either side of the *dfrbs* and classified them according to the antimicrobial to which they confer resistance (Figure 3.2.a,b). The most prevalent phenotype was aminoglycoside resistance (80% of all sequences), followed by beta-lactamase genes (60%). Surprisingly, although TMP is often prescribed in combination with sulfonamide⁴², only 57% of sequences included sulfonamide-resistance genes. Resistance to metals, phenicol, fluoroquinolones, rifampicin, macrolides, lincosamides, streptogramins, efflux genes and tetracycline were observed at a lower frequency.

3.4 Discussion

The overwhelming majority of studies on clinical resistance to trimethoprim have focused exclusively on DfrAs. Recently, Sánchez-Osuna *et al.* reported two mechanisms of DfrA evolution⁹. One mechanism involves the mutation and mobilization of trimethoprim-sensitive *FolA* genes (ex. *A. baumannii folA* mutated to *dfrA39* and *dfrA40*); the second relies on the mobilization of intrinsically trimethoprim-resistant *folA* genes. Trimethoprim resistance through DfrAs evolves readily, explaining the large number of DfrA genes: the recent addition of four members to the DfrA family has brought it to nearly 40 members⁹.

Unlike DfrAs, the evolutionary origin of DfrBs has not been investigated. Little is known about these peculiar homotetrameric enzymes, including their prevalence and emergence in clinical and environmental samples²⁷. We examine, for the first time, the genomic context of *dfrbs* by analyzing publicly available sequences containing *dfrbs*. By reporting the microorganisms that harbor them and determining whether they occur in the context of genetic mobility and/or resistance to multiple antibiotics, we provide insight into the risk they represent for public health.

Using a query-set consisting of the eight previously known *dfrb* genes and two newly identified and confirmed TMP-resistant *dfrb* genes, we identified 61 different genomic sequences containing *dfrbs*. The country of origin of each sample illustrates that *dfrbs* are dispersed worldwide. The vast majority of *dfrbs* (74%) for which the source of isolation is known are related to human activities, whereas only one sequence comes from an environmental sample and 25% are from unknown sources. This observation could well reflect sampling biases due to overrepresentation of studies related to human activities relative to environmental studies in genomic databases. In Canada, TMP and sulfonamides are the fourth most highly prescribed antimicrobials for animals, representing 57,865 kg in 2018⁴³, justifying the importance of increasing genomic analyses of animal samples to determine the prevalence of *dfrbs* in all relevant contexts. Amongst the chromosomal sequences we identified, 93% contain virulence genes and include at least one MGE near the *dfrb* gene. These combined criteria define the highest risk that antimicrobial resistance genes can present³². The remaining sequences containing *dfrbs* include at least one of these two criteria. All plasmid-borne *dfrbs* are near a MGE, allowing them to spread easily among bacteria.

Since their discovery, DfrBs have been considered to be solely plasmid-borne^{44,45}. Indeed, *dfrb1* was first observed in an *E. coli* strain where it is plasmid-borne, leading to the incorrect assumption that *dfrbs* are always plasmidic³⁸. Nevertheless, we observed not only a few exceptions to this long-standing conjecture, but rather that only 29% of *dfrb*-containing sequences in our sample are plasmidic. The *Enterbacterales* species harbor *dfrbs* on plasmids, while *Aeromonadales* and *Pseudomonadales* species harbor *dfrbs* on their chromosome (Figure 3.1).

It is interesting to note that the only two *dfrb* genes that were not found near MGE are in chromosomal sequences (Figure 3.1). This suggests that *dfrbs* might have mobilized from a chromosome to a plasmid, and not have originated from plasmids. A more thorough examination of *dfrbs* and their mobilization context, both in plasmids and chromosomes, will allow retracing of the early events of *dfrb* mobilization.

DfrB1 and DfrB2 were discovered in the 1970s, subsequent to the introduction of TMP. DfrB1, also named R67, dfrII and *dfr2a*, has been extensively characterized for its structure, assembly and catalytic activity^{14,16,17,46-49}. Only recently have other members of the family been characterized¹⁹ or even reported, such that it could have been thought that DfrB1 is the most widespread among DfrBs. However, *dfrb1* was identified in only 16% of the 61 sequences identified here. Unexpectedly, *dfrb5* was identified in 38% of the sequences, in various genomic contexts and geographical locations, suggesting that it is more broadly disseminated than *dfrb1*. In addition, 23% of sequences contained *dfrb2*, 13% contained *dfrb3* while 8% contained *dfrb4*. The *dfrb10* gene, reported here for the first time, was identified in one sequence. The *dfrb6*, *dfrb7*, *dfrb9* or *dfrb11* genes were not identified using our search criteria for genomic segments.

Although the sample sets included in the databases we searched represent only a partial picture of gene dissemination, some members of the DfrB family are clearly more prevalent while others may not have yet emerged. Interestingly, the most prevalent *dfrb5* is closely related to the first-reported *dfrb1* (Table S3.1). Further investigations will be required to determine whether early events of *dfrb* mobilization in pathogenic bacteria involves either of these two genes. Close surveillance of the prevalence of these genes is needed to evaluate the spread of this family of genes.

All but one *dfrb* (NZ_CM002277.1) were found near at least one antibiotic resistance gene (ARG), the majority (83%) being in proximity (less than 10kb) to 3 other ARG. Among these, the *sul1* sulfonamide-resistant gene is present in the vast majority of clinically relevant integrons⁵⁰. We previously reported the identification of *dfrb4* in a clinical class 1 integron within the *dfrb4/qacEΔ1/sul1* cassette, flanked by other resistance genes²⁷. Since TMP is often prescribed in combination with sulfamethoxazole, it is noteworthy that only 57% of the *dfrbs* identified in integrons in this study were colocalized with *sul1*. The *dfrbs* were found in class 1 integrons as defined by the presence of a class 1 integrase. Multiple ARG were observed within the same cassettes as the *dfrbs* (Figure 3.2.a). Expression of some of these ARG (e.g. β-lactam resistance *vim-2* and *oxa-10*, aminoglycoside resistance *aadA1*, rifampin resistance *arr-2*) in other class 1 integrons has previously been reported⁵¹⁻⁵³. Although this does not demonstrate gene expression in these genomic contexts, it is consistent with the hypothesis that the ARG as well as the *dfrbs* are expressed in the class 1 integrons identified here.

No *dfras* were found in proximity to the *dfrbs*; this is expected since these genes produce the same phenotype. Nonetheless, duplication of *dfrb5* was observed in the integron of one genome

(NZ_CP031449.2), where three copies of *dfrb5* were observed in the same integron (Figure 3.2.a). In addition, duplication of similar integrons containing *dfrb5* in the same sequence was observed in one plasmid and four genomes, all from *P. aeruginosa*. Considering the incomplete nature of the sequences we analyzed, it is possible that a greater number of amplification events could be identified upon analysis of longer sequence segments.

Given the importance of TMP both in the clinic and with livestock, it is critical to monitor the emergence of resistance to this antimicrobial. TMP resistance has generally been associated with DfrAs. Here, we have demonstrated that monitoring the emergence and prevalence of DfrBs will provide important insights into global TMP resistance and thus contribute to policy making to contain the spread of antimicrobial resistance.

3.5 Materials and Methods

3.5.1 Identification of putative type B dihydrofolate reductases

The six DfrB sequences that were previously functionally characterized (DfrB1 –DfrB5 and DfrB7)¹⁹ were used to create a profile hidden Markov models (HMM) with HMMER version 3.3 (<http://hmmer.org/>). This profile was used as a query against the UniProtKB/TrEMBL database (Apr-22, 2020 release, 184,998,855 sequences)⁵⁴. Hits with E-value lower than 1e-40 were considered and compared to known DfrB sequences. Predicted sequences having protein sequence identity lower than 95% relative to any known DfrB sequence were considered as new genes.

3.5.2 Subcloning of *dfrb10* and *dfrb11* genes

The genetic sequences of *dfrb10* and *dfrb11* were obtained in pUC57 (BioBasic) according to the Genbank accession numbers in Table 3.1. The N-terminally His₆-tagged ORF sequences of *dfrb10* and *dfrb11* were subcloned into pET24 (Qiagen) upstream of the lactose operon repressor, following the lac operator sequence, using the *Nde*I and *Hind*III restriction sites. Both genes were amplified by PCR using the same forward primer 5'-GAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGAGAGGATCTCACCATCAC-3' (*Nde*I site in bold) and a reverse primer that differs at one base (underlined) to maintain the native stop codon, *dfrb10*: 5'-GGTGGTGCTCGAGTGCGGCCGCAAGCTTTTAGGCCACGCG-3'; *dfrb11*: 5'-GGTGGTGCTCGAGTGCGGCCGCAAGCTTTCAGGCCACGCG-3' (*Hind*III site in bold). Phusion HF polymerase (ThermoFisher) was used according to the manufacturer's protocol, using 55°C as the annealing temperature. Amplified genes, as well as pET24, were digested with *Hind*III (NEB) for 14 h and *Nde*I (NEB) for 2 h at 37°C, followed by enzyme inactivation for 20 min at 80°C. They were gel-extracted using the Monarch DNA gel extraction kit (NEB) and purified using the DNA Cleanup kit (NEB). Inserts were ligated into digested pET24 using a DNA ligation kit (Takara) according to the manufacturer's instructions.

Briefly, the digested gene and pET24 vector were incubated at 16°C for 3 h in Takara solution I. The ligation products were transformed into CaCl₂-competent *E. coli* DH5 α prepared by the method of Inoue⁵⁵. The DNA sequences of *dfrB10*-pET24 and *dfrB11*-pET24 were confirmed by DNA Sanger sequencing (Genome Quebec platform at Sainte-Justine Hospital). The final constructs yield N-terminally, His₆-tagged DfrB proteins. His₆-DfrB3 in pET24 was previously reported¹⁹. The negative control cTEM-19m, with an expressible β -lactamase insert instead of a *dfrb* insert, was previously described⁵⁶.

3.5.3 Minimal Inhibitory Concentration (MIC)

MICs were determined according to Wiegand *et al*⁵⁷ using the broth microdilution method. Briefly, *E. coli* BL21(DE3) cells expressing His₆-DfrB3 (positive control), His₆-DfrB10, His₆-DfrB11 or cTEM-19m (negative control) were propagated overnight in Luria-Bertani (LB) media with 50 μ g/mL kanamycin. In 96-well plates, an inoculum of 10⁵ colony forming units (cfu) was inoculated in LB medium, with 0.1 mM IPTG (ThermoFisher) and TMP (Sigma) in 2-fold concentration steps up to 600 μ g/mL; the latter is the highest concentration of TMP soluble in 5% methanol. MICs were determined in triplicate.

3.5.4 Download of genomes

The sequences of *dfrb1-dfrb7*, *dfrb9* and the newly identified *dfrb10* and *dfrb11* were used as queries for blastn 2.10.0 searches against four genomic databases (performed on the 2020.07.04): RefSeq (bacterial sequences), GenBank (bacterial sequences) and the Microbial Complete Genomes and Complete Plasmids databases found at <https://blast.ncbi.nlm.nih.gov/Blast.cgi>⁵⁸⁻⁶⁰. Genomes containing at least one query sequence and having a sequence length of at least 10 kb both upstream and downstream of the *dfrb* sequences were collated. In total, 110 sequences were identified and served for analysis.

3.5.5 Protein database constructions

The following protein databases were downloaded on the 2020.07.07: Integrase, IntI1 and sul1 databases from the I-VIP pipeline⁶¹, ARG-ANNOT²⁵, ICEberg 2.0⁶², CARD²³, BacMet⁶³ and UniProtKB/Swiss-Prot⁶⁴. These databases were merged and redundant sequences were removed. Two genes, coding for 78 and 97 amino-acid products, were both named DfrB1; the shorter gene version was kept to match the consensus length of all other members of the family.

3.5.6 Annotation

The 110 sequences of 20 kb (10 kb upstream and downstream of a *dfrb*) were used as query sequences against the blast database for a blastx 2.10.0 search with the parameters of E-value lower than 1e-10 and culling_limit of 1. Hits with coverage of \geq 80% and protein identity of \geq 60% were kept. Where multiple sequences from a same NCBI BioProject presented the same annotation, all but one were removed from the dataset. In total, 61 sequences served for analysis.

3.5.7 Classification of sequences as chromosomal or plasmidic

The 110 sequences were classified as chromosomal or plasmidic using PlasFlow 1.1 with a threshold of 0.65⁴¹.

3.5.8 Identification of pathogenic hosts

For each sequence, a blastx 2.10.0 analysis was carried against virulence factor protein sequences from the core dataset of VFDB (last update on 07/17/2020)⁶⁵ using an e-value cutoff of 1e-15. Hits were filtered using an identity and coverage threshold of 60%. Sequences with one or more hits were labelled pathogenic.

3.5.9 Phylogenetic tree

The phylogenetic tree was constructed with NGPhylogeny.fr⁶⁶ using the host species' 16S rRNA from the NCBI reference genome. The tools MUSCLE (with the Neighbour joining option), Noisy, PHyML + SMS, and Newick were used for tree construction. The tree was annotated with iTOL, where the GenBank accession of each sequence is displayed⁶⁷.

3.6 Author Contributions

C.L.S.D. conceptualization, methodology, data analysis, writing—original draft preparation; S.S.D. data analysis; J.N.P. conceptualization, writing—review and editing, funding acquisition. All authors have read and agreed to the published version of the manuscript.

3.7 Funding

This research was funded by NSERC grant number RGPIN-2018-04686. C.L.S.D. and S.S.D. are grateful to NSERC, FQRNT and Université de Montréal for scholarships.

3.8 Acknowledgments

We thank Janine N. Copp for her insights and fruitful discussion, as well as Lorea Alejaldre and Étienne Lavallée for proofreading. We also thank William Huynh and Andrew McArthur at CARD and Daniel Haft at NCBI for their help navigating various databases.

3.9 Conflict of Interest

The authors declare no conflict of interest.

3.10 References

- (1) World Health Organization. Critically Important Antimicrobials for Human Medicine, 6th Revision, 2018.
- (2) Stone, S. R.; Morrison, J. F. Mechanism of Inhibition of Dihydrofolate Reductases from Bacterial and Vertebrate Sources by Various Classes of Folate Analogues. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* **1986**, *869* (3), 275–285. [https://doi.org/10.1016/0167-4838\(86\)90067-1](https://doi.org/10.1016/0167-4838(86)90067-1).

- (3) Eliopoulos, G. M.; Huovinen, P. Resistance to Trimethoprim-Sulfamethoxazole. *Clin. Infect. Dis.* **2001**, *32* (11), 1608–1614. <https://doi.org/10.1086/320532>.
- (4) Caron, F.; Wehrle, V.; Etienne, M. The Comeback of Trimethoprim in France. *Médecine Mal. Infect.* **2017**, *47* (4), 253–260. <https://doi.org/10.1016/j.medmal.2016.12.001>.
- (5) World Health Organization. WHO Report on Surveillance of Antibiotic Consumption: 2016-2018 Early Implementation, 2018.
- (6) Hsia, Y.; Lee, B. R.; Versporten, A.; Yang, Y.; Bielicki, J.; Jackson, C.; Newland, J.; Goossens, H.; Magrini, N.; Sharland, M.; GARPEC and Global-PPS networks. Use of the WHO Access, Watch, and Reserve Classification to Define Patterns of Hospital Antibiotic Use (AWaRe): An Analysis of Paediatric Survey Data from 56 Countries. *Lancet Glob. Health* **2019**, *7* (7), e861–e871. [https://doi.org/10.1016/S2214-109X\(19\)30071-3](https://doi.org/10.1016/S2214-109X(19)30071-3).
- (7) Cuong, N. V.; Padungtod, P.; Thwaites, G.; Carrique-Mas, J. J. Antimicrobial Usage in Animal Production: A Review of the Literature with a Focus on Low- and Middle-Income Countries. *Antibiot. Basel Switz.* **2018**, *7* (3). <https://doi.org/10.3390/antibiotics7030075>.
- (8) World Health Organization; Department of Food Safety and Zoonoses; World Health Organization. *WHO Guidelines on Use of Medically Important Antimicrobials in Food-Producing Animals.*; 2017.
- (9) Sánchez-Osuna, M.; Cortés, P.; Llagostera, M.; Barbé, J.; Erill, I. Exploration into the Origins and Mobilization of Di-Hydrofolate Reductase Genes and the Emergence of Clinical Resistance to Trimethoprim. *Microb. Genomics* **2020**, *6* (11). <https://doi.org/10.1099/mgen.0.000440>.
- (10) Howell, E. E. Searching Sequence Space: Two Different Approaches to Dihydrofolate Reductase Catalysis. *ChemBioChem* **2005**, *6* (4), 590–600. <https://doi.org/10.1002/cbic.200400237>.
- (11) Faltyn, M.; Alcock, B.; McArthur, A. Evolution and Nomenclature of the Trimethoprim Resistant Dihydrofolate (Dfr) Reductases. **2019**. <https://doi.org/10.20944/preprints201905.0137.v1>.
- (12) Kim, Y. B. Improved Trimethoprim-Resistance Cassette for Prokaryotic Selections. *J. Biosci. Bioeng.* **2009**, *108* (5), 441–445. <https://doi.org/10.1016/j.jbiosc.2009.05.015>.
- (13) Pattishall, K. H.; Acar, J.; Burchall, J. J.; Goldstein, F. W.; Harvey, R. J. Two Distinct Types of Trimethoprim-Resistant Dihydrofolate Reductase Specified by R-Plasmids of Different Compatibility Groups. *J. Biol. Chem.* **1977**, *252* (7), 2319–2323.
- (14) Krahn, J. M.; Jackson, M. R.; DeRose, E. F.; Howell, E. E.; London, R. E. Crystal Structure of a Type II Dihydrofolate Reductase Catalytic Ternary Complex[†]. *Biochemistry* **2007**, *46* (51), 14878–14888. <https://doi.org/10.1021/bi701532r>.
- (15) Bhojane, P. P.; Duff, M. R.; Bafna, K.; Agarwal, P.; Stanley, C.; Howell, E. E. Small Angle Neutron Scattering Studies of R67 Dihydrofolate Reductase, a Tetrameric Protein with Intrinsically Disordered N-Termini. *Biochemistry* **2017**, *56* (44), 5886–5899. <https://doi.org/10.1021/acs.biochem.7b00822>.
- (16) Ebert, M. C. C. J. C.; Morley, K. L.; Volpato, J. P.; Schmitzer, A. R.; Pelletier, J. N. Asymmetric Mutations in the Tetrameric R67 Dihydrofolate Reductase Reveal High Tolerance to Active-Site Substitutions: Asymmetric Mutations in R67 Dihydrofolate Reductase. *Protein Sci.* **2015**, *24* (4), 495–507. <https://doi.org/10.1002/pro.2602>.
- (17) Schmitzer, A. R.; Lépine, F.; Pelletier, J. N. Combinatorial Exploration of the Catalytic Site of a Drug-Resistant Dihydrofolate Reductase: Creating Alternative Functional Configurations. *Protein Eng. Des. Sel.* **2004**, *17* (11), 809–819. <https://doi.org/10.1093/protein/gzh090>.
- (18) Martinez, M. A.; Pezo, V.; Marlière, P.; Wain-Hobson, S. Exploring the Functional Robustness of an Enzyme by in Vitro Evolution. *EMBO J.* **1996**, *15* (6), 1203–1210.

- (19) Toulouse, J. L.; Shi, G.; Lemay-St-Denis, C.; Ebert, M. C. C. J. C.; Deon, D.; Gagnon, M.; Ruediger, E.; Saint-Jacques, K.; Forge, D.; Vanden Eynde, J. J.; Marinier, A.; Ji, X.; Pelletier, J. N. Dual-Target Inhibitors of the Folate Pathway Inhibit Intrinsically Trimethoprim-Resistant DfrB Dihydrofolate Reductases. *ACS Med. Chem. Lett.* **2020**, acsmedchemlett.0c00393. <https://doi.org/10.1021/acsmchemlett.0c00393>.
- (20) Toulouse, J. L.; Yachnin, B. J.; Ruediger, E. H.; Deon, D.; Gagnon, M.; Saint-Jacques, K.; Ebert, M. C. C. J. C.; Forge, D.; Bastien, D.; Colin, D. Y.; Vanden Eynde, J. J.; Marinier, A.; Berghuis, A. M.; Pelletier, J. N. Structure-Based Design of Dimeric Bisbenzimidazole Inhibitors to an Emergent Trimethoprim-Resistant Type II Dihydrofolate Reductase Guides the Design of Monomeric Analogues. *ACS Omega* **2019**, *4* (6), 10056–10069. <https://doi.org/10.1021/acsomega.9b00640>.
- (21) Toulouse, J. L.; Abraham, S. M. J.; Kadnikova, N.; Bastien, D.; Gauchot, V.; Schmitzer, A. R.; Pelletier, J. N. Investigation of Classical Organic and Ionic Liquid Cosolvents for Early-Stage Screening in Fragment-Based Inhibitor Design with Unrelated Bacterial and Human Dihydrofolate Reductases. *Assay Drug Dev. Technol.* **2017**, *15* (4), 141–153. <https://doi.org/10.1089/adt.2016.768>.
- (22) Bastien, D.; Ebert, M. C. C. J. C.; Forge, D.; Toulouse, J.; Kadnikova, N.; Perron, F.; Mayence, A.; Huang, T. L.; Vanden Eynde, J. J.; Pelletier, J. N. Fragment-Based Design of Symmetrical Bis-Benzimidazoles as Selective Inhibitors of the Trimethoprim-Resistant, Type II R67 Dihydrofolate Reductase. *J. Med. Chem.* **2012**, *55* (7), 3182–3192. <https://doi.org/10.1021/jm201645r>.
- (23) Alcock, B. P.; Raphenya, A. R.; Lau, T. T. Y.; Tsang, K. K.; Bouchard, M.; Edalatmand, A.; Huynh, W.; Nguyen, A.-L. V.; Cheng, A. A.; Liu, S.; Min, S. Y.; Miroshnichenko, A.; Tran, H.-K.; Werfalli, R. E.; Nasir, J. A.; Oloni, M.; Speicher, D. J.; Florescu, A.; Singh, B.; Faltyn, M.; Hernandez-Koutoucheva, A.; Sharma, A. N.; Bordeleau, E.; Pawlowski, A. C.; Zubyk, H. L.; Dooley, D.; Griffiths, E.; Maguire, F.; Winsor, G. L.; Beiko, R. G.; Brinkman, F. S. L.; Hsiao, W. W. L.; Domselaar, G. V.; McArthur, A. G. CARD 2020: Antibiotic Resistome Surveillance with the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **2019**, gkz935. <https://doi.org/10.1093/nar/gkz935>.
- (24) Liu, B.; Pop, M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **2009**, *37* (Database), D443–D447. <https://doi.org/10.1093/nar/gkn656>.
- (25) Gupta, S. K.; Padmanabhan, B. R.; Diene, S. M.; Lopez-Rojas, R.; Kempf, M.; Landraud, L.; Rolain, J.-M. ARG-ANNOT, a New Bioinformatic Tool to Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrob. Agents Chemother.* **2014**, *58* (1), 212–220. <https://doi.org/10.1128/AAC.01310-13>.
- (26) Grape, M.; Farra, A.; Kronvall, G.; Sundström, L. Integrons and Gene Cassettes in Clinical Isolates of Co-Trimoxazole-Resistant Gram-Negative Bacteria. *Clin. Microbiol. Infect.* **2005**, *11* (3), 185–192. <https://doi.org/10.1111/j.1469-0691.2004.01059.x>.
- (27) Toulouse, J. L.; Edens, T. J.; Alejaldre, L.; Manges, A. R.; Pelletier, J. N. Integron-Associated DfrB4, a Previously Uncharacterized Member of the Trimethoprim-Resistant Dihydrofolate Reductase B Family, Is a Clinically Identified Emergent Source of Antibiotic Resistance. *Antimicrob. Agents Chemother.* **2017**, *61* (5), e02665-16, /aac/61/5/e02665-16.atom. <https://doi.org/10.1128/AAC.02665-16>.
- (28) Dinger, M. E.; Pang, K. C.; Mercer, T. R.; Mattick, J. S. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLoS Comput. Biol.* **2008**, *4* (11), e1000176. <https://doi.org/10.1371/journal.pcbi.1000176>.
- (29) Hyatt, D.; Chen, G.-L.; LoCascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics* **2010**, *11* (1), 119. <https://doi.org/10.1186/1471-2105-11-119>.

- (30) Kadlec, K.; Kehrenberg, C.; Schwarz, S. Molecular Basis of Resistance to Trimethoprim, Chloramphenicol and Sulphonamides in *Bordetella Bronchiseptica*. *J. Antimicrob. Chemother.* **2005**, *56* (3), 485–490. <https://doi.org/10.1093/jac/dki262>.
- (31) Grape, M.; Motakefi, A.; Pavuluri, S.; Kahlmeter, G. Standard and Real-Time Multiplex PCR Methods for Detection of Trimethoprim Resistance Dfr Genes in Large Collections of Bacteria. *Clin. Microbiol. Infect.* **2007**, *13* (11), 1112–1118. <https://doi.org/10.1111/j.1469-0691.2007.01807.x>.
- (32) Martínez, J. L.; Coque, T. M.; Baquero, F. What Is a Resistance Gene? Ranking Risk in Resistomes. *Nat. Rev. Microbiol.* **2015**, *13* (2), 116–123. <https://doi.org/10.1038/nrmicro3399>.
- (33) Jaillard, M.; van Belkum, A.; Cady, K. C.; Creely, D.; Shortridge, D.; Blanc, B.; Barbu, E. M.; Dunne, W. M.; Zambardi, G.; Enright, M.; Mugnier, N.; Le Priol, C.; Schicklin, S.; Guigon, G.; Veyrieras, J.-B. Correlation between Phenotypic Antibiotic Susceptibility and the Resistome in *Pseudomonas Aeruginosa*. *Int. J. Antimicrob. Agents* **2017**, *50* (2), 210–218. <https://doi.org/10.1016/j.ijantimicag.2017.02.026>.
- (34) Azam, M. W.; Khan, A. U. Updates on the Pathogenicity Status of *Pseudomonas Aeruginosa*. *Drug Discov. Today* **2019**, *24* (1), 350–359. <https://doi.org/10.1016/j.drudis.2018.07.003>.
- (35) Sun, F.; Zhou, D.; Wang, Q.; Feng, J.; Feng, W.; Luo, W.; Liu, Y.; Qiu, X.; Yin, Z.; Xia, P. Genetic Characterization of a Novel *Bla*_{DIM-2}-Carrying Megaplasmid P12969-DIM from Clinical *Pseudomonas Putida*. *J. Antimicrob. Chemother.* **2016**, *71* (4), 909–912. <https://doi.org/10.1093/jac/dkv426>.
- (36) Hernsdorf, A. W.; Amano, Y.; Miyakawa, K.; Ise, K.; Suzuki, Y.; Anantharaman, K.; Probst, A.; Burstein, D.; Thomas, B. C.; Banfield, J. F. Potential for Microbial H₂ and Metal Transformations Associated with Novel Bacteria and Archaea in Deep Terrestrial Subsurface Sediments. *ISME J.* **2017**, *11* (8), 1915–1929. <https://doi.org/10.1038/ismej.2017.39>.
- (37) Workentine, M. L.; Surette, M. G.; Bernier, S. P. Draft Genome Sequence of *Burkholderia Dolosa* PC543 Isolated from Cystic Fibrosis Airways. *Genome Announc.* **2014**, *2* (1). <https://doi.org/10.1128/genomeA.00043-14>.
- (38) Stone, D.; Smith, S. L. The Amino Acid Sequence of the Trimethoprim-Resistant Dihydrofolate Reductase Specified in *Escherichia Coli* by R-Plasmid R67. *J. Biol. Chem.* **1979**, *254* (21), 10857–10861.
- (39) Amyes, S. G. B.; Smith, J. T. R-Factor Trimethoprim Resistance Mechanism: An Insusceptible Target Site. *Biochem. Biophys. Res. Commun.* **1974**, *58* (2), 412–418. [https://doi.org/10.1016/0006-291X\(74\)90380-5](https://doi.org/10.1016/0006-291X(74)90380-5).
- (40) San Millan, A. Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context. *Trends Microbiol.* **2018**, *26* (12), 978–985. <https://doi.org/10.1016/j.tim.2018.06.007>.
- (41) Krawczyk, P. S.; Lipinski, L.; Dziembowski, A. PlasFlow: Predicting Plasmid Sequences in Metagenomic Data Using Genome Signatures. *Nucleic Acids Res.* **2018**, *46* (6), e35–e35. <https://doi.org/10.1093/nar/gkx1321>.
- (42) Masters, P. A.; O’Byrne, T. A.; Zurlo, J.; Miller, D. Q.; Joshi, N. Trimethoprim-Sulfamethoxazole Revisited. *Arch. Intern. Med.* **2003**, *163* (4), 402–410. <https://doi.org/10.1001/archinte.163.4.402>.
- (43) Public Health Agency of Canada. Canadian Antimicrobial Resistance Surveillance System Report, 2020. <https://www.canada.ca/content/dam/hc-sc/documents/services/drugs-health-products/canadian-antimicrobial-resistance-surveillance-system-2020-report/CARSS-2020-report-2020-eng.pdf>.

- (44) Swift, G.; McCarthy, B. J.; Heffron, F. DNA Sequence of a Plasmid-Encoded Dihydrofolate Reductase. *Mol. Gen. Genet. MGG* **1981**, *181* (4), 441–447. <https://doi.org/10.1007/BF00428733>.
- (45) Brisson, N.; Hohn, T. Nucleotide Sequence of the Dihydrofolate-Reductase Gene Borne by the Plasmid R67 and Conferring Methotrexate Resistance. *Gene* **1984**, *28* (2), 271–274. [https://doi.org/10.1016/0378-1119\(84\)90266-x](https://doi.org/10.1016/0378-1119(84)90266-x).
- (46) Narayana, N.; Matthews, D. A.; Howell, E. E.; Xuong, N. A Plasmid-Encoded Dihydrofolate Reductase from Trimethoprim-Resistant Bacteria Has a Novel D2-Symmetric Active Site. *Nat. Struct. Mol. Biol.* **1995**, *2* (11), 1018–1025. <https://doi.org/10.1038/nsb1195-1018>.
- (47) Park, H.; Zhuang, P.; Nichols, R.; Howell, E. E. Mechanistic Studies of R67 Dihydrofolate Reductase. *J. Biol. Chem.* **1997**, *272* (4), 2252–2258. <https://doi.org/10.1074/jbc.272.4.2252>.
- (48) West, F. W.; Seo, H.-S.; Bradrick, T. D.; Howell, E. E. Effects of Single-Tryptophan Mutations on R67 Dihydrofolate Reductase †. *Biochemistry* **2000**, *39* (13), 3678–3689. <https://doi.org/10.1021/bi992195x>.
- (49) Kamath, G.; Howell, E. E.; Agarwal, P. K. The Tail Wagging the Dog: Insights into Catalysis in R67 Dihydrofolate Reductase. *Biochemistry* **2010**, *49* (42), 9078–9088. <https://doi.org/10.1021/bi1007222>.
- (50) Gillings, M. R. Class 1 Integrons as Invasive Species. *Curr. Opin. Microbiol.* **2017**, *38*, 10–15. <https://doi.org/10.1016/j.mib.2017.03.002>.
- (51) Santos, C.; Caetano, T.; Ferreira, S.; Mendo, S. Tn5090-like Class 1 Integron Carrying BlaVIM-2 in a Pseudomonas Putida Strain from Portugal. *Clin. Microbiol. Infect.* **2010**, *16* (10), 1558–1561. <https://doi.org/10.1111/j.1469-0691.2010.03165.x>.
- (52) Chiu, C.-H.; Lee, H.-Y.; Tseng, L.-Y.; Chen, C.-L.; Chia, J.-H.; Su, L.-H.; Liu, S.-Y. Mechanisms of Resistance to Ciprofloxacin, Ampicillin/Sulbactam and Imipenem in Acinetobacter Baumannii Clinical Isolates in Taiwan. *Int. J. Antimicrob. Agents* **2010**, *35* (4), 382–386. <https://doi.org/10.1016/j.ijantimicag.2009.12.009>.
- (53) Naas, T.; Mikami, Y.; Imai, T.; Poirel, L.; Nordmann, P. Characterization of In53, a Class 1 Plasmid- and Composite Transposon-Located Integron of Escherichia Coli Which Carries an Unusual Array of Gene Cassettes. *J. Bacteriol.* **2001**, *183* (1), 235–249. <https://doi.org/10.1128/JB.183.1.235-249.2001>.
- (54) Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.-C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003. *Nucleic Acids Res.* **2003**, *31* (1), 365–370. <https://doi.org/10.1093/nar/gkg095>.
- (55) Sambrook, J.; Russell, D. W. The Inoue Method for Preparation and Transformation of Competent *E. Coli*: “Ultra-Competent” Cells. *Cold Spring Harb. Protoc.* **2006**, *2006* (1), pdb.prot3944. <https://doi.org/10.1101/pdb.prot3944>.
- (56) Gobeil, S. M. C.; Gagné, D.; Doucet, N.; Pelletier, J. N. 15N, 13C and 1H Backbone Resonance Assignments of an Artificially Engineered TEM-1/PSE-4 Class A β -Lactamase Chimera and Its Deconvoluted Mutant. *Biomol. NMR Assign.* **2016**, *10* (1), 93–99. <https://doi.org/10.1007/s12104-015-9645-8>.
- (57) Wiegand, I.; Hilpert, K.; Hancock, R. E. W. Agar and Broth Dilution Methods to Determine the Minimal Inhibitory Concentration (MIC) of Antimicrobial Substances. *Nat. Protoc.* **2008**, *3* (2), 163–175. <https://doi.org/10.1038/nprot.2007.521>.

- (58) Sayers, E. W.; Agarwala, R.; Bolton, E. E.; Brister, J. R.; Canese, K.; Clark, K.; Connor, R.; Fiorini, N.; Funk, K.; Hefferon, T.; Holmes, J. B.; Kim, S.; Kimchi, A.; Kitts, P. A.; Lathrop, S.; Lu, Z.; Madden, T. L.; Marchler-Bauer, A.; Phan, L.; Schneider, V. A.; Schoch, C. L.; Pruitt, K. D.; Ostell, J. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2019**, *47* (D1), D23–D28. <https://doi.org/10.1093/nar/gky1069>.
- (59) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Res.* **2005**, *33* (Database issue), D501–504. <https://doi.org/10.1093/nar/gki025>.
- (60) Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res.* **2017**, *45* (D1), D37–D42. <https://doi.org/10.1093/nar/gkw1070>.
- (61) Zhang, A. N.; Li, L.-G.; Ma, L.; Gillings, M. R.; Tiedje, J. M.; Zhang, T. Conserved Phylogenetic Distribution and Limited Antibiotic Resistance of Class 1 Integrons Revealed by Assessing the Bacterial Genome and Plasmid Collection. *Microbiome* **2018**, *6* (1), 130. <https://doi.org/10.1186/s40168-018-0516-2>.
- (62) Liu, M.; Li, X.; Xie, Y.; Bi, D.; Sun, J.; Li, J.; Tai, C.; Deng, Z.; Ou, H.-Y. ICEberg 2.0: An Updated Database of Bacterial Integrative and Conjugative Elements. *Nucleic Acids Res.* **2019**, *47* (D1), D660–D665. <https://doi.org/10.1093/nar/gky1123>.
- (63) Pal, C.; Bengtsson-Palme, J.; Rensing, C.; Kristiansson, E.; Larsson, D. G. J. BacMet: Antibacterial Biocide and Metal Resistance Genes Database. *Nucleic Acids Res.* **2014**, *42* (D1), D737–D743. <https://doi.org/10.1093/nar/gkt1252>.
- (64) Bairoch, A.; Boeckmann, B. The SWISS-PROT Protein Sequence Data Bank. *Nucleic Acids Res.* **1991**, *19 Suppl*, 2247–2249. <https://doi.org/10.1093/nar/19.suppl.2247>.
- (65) Liu, B.; Zheng, D.; Jin, Q.; Chen, L.; Yang, J. VFDB 2019: A Comparative Pathogenomic Platform with an Interactive Web Interface. *Nucleic Acids Res.* **2019**, *47* (D1), D687–D692. <https://doi.org/10.1093/nar/gky1080>.
- (66) Lemoine, F.; Correia, D.; Lefort, V.; Doppelt-Azeroual, O.; Mareuil, F.; Cohen-Boulakia, S.; Gascuel, O. NGPhylogeny.Fr: New Generation Phylogenetic Services for Non-Specialists. *Nucleic Acids Res.* **2019**, *47* (W1), W260–W265. <https://doi.org/10.1093/nar/gkz303>.
- (67) Letunic, I.; Bork, P. Interactive Tree of Life (ITOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees. *Nucleic Acids Res.* **2016**, *44* (W1), W242–245. <https://doi.org/10.1093/nar/gkw290>.

3.11 Supplementary material

Table S3.1. Protein sequence identity of the ten members of the DfrB family ^a

	DfrB1	DfrB2	DfrB3	DfrB4	DfrB5	DfrB6	DfrB7	DfrB9	DfrB10	DfrB11
DfrB1	100%	78%	78%	77%	88%	87%	88%	77%	82%	78%
DfrB2	17	100%	86%	74%	79%	82%	78%	84%	92%	88%
DfrB3	17	11	100%	79%	82%	83%	79%	85%	92%	90%
DfrB4	18	20	16	100%	77%	77%	76%	74%	77%	79%
DfrB5	9	16	14	18	100%	91%	94%	77%	83%	81%
DfrB6	10	14	13	18	7	100%	91%	79%	86%	83%
DfrB7	9	17	16	19	5	7	100%	76%	82%	81%
DfrB9	18	12	12	20	18	16	19	100%	88%	86%
DfrB10	14	6	6	16	13	11	14	9	100%	92%
DfrB11	17	9	8	18	15	13	15	11	6	100%

^a Values in percentage correspond to protein sequence identity values; values in grey represent the number of substitutions between sequences.

Table S3.2. DfrB sequences

Gene	Genbank ID	Gene sequence	Protein sequence
<i>dfrb1</i>	KAB1871659.1	ATGGAACGAAGTAGCAATGAAGTCAGTAATCCAGTTGCTGGCAATTTT GTATTCCCATCGAACGCCACGTTTGGTATGGGAGATCGCGTGCACAAG AAATCCGGCGCCGCTGGCAAGGTCAGATTGTCGGGTGGTACTGCAC AAATTTGACCCCGAAGGCTACGCCGTCGAGTCTGAGGCTACCCAGG CTCAGTACAGATTTATCCTGTTGCGGCGCTTGAACGCATCAACTGA	MERSSNEVSNPVAGNFV FPSNATFGMGDRVRKKS GAAWQQQIVGWYCTNL TPEGYAVESEAHPGSVQI YPVAALERIN
<i>dfrb2</i>	FAA00064.1	ATGGGTCAAAGTAGCGATGAAGCCAACGCTCCCGTTGCAGGGCAGTTT GCGCTTCCCCTGAGTGCCACCTTTGGCTTAGGGGATCGCGTACGCAAG AAATCTGGTGCCGCTTGGCAGGGTCAAGTCGTCGGTGGTATTGCACA AAACTCACTCCTGAAGGCTATGCGGTGCGAGTCCGAATCCCACCCAGGC TCAGTGCAAATTTATCCTGTGGCTGCACTTGAACGTGTGGCCTAA	MGQSSDEANAPVAGQF ALPLSATFGLGDRVRKKS GAAWQQQVVGWYCTKL TPEGYAVESHPGVSQI YPVAALERVA
<i>dfrb3</i>	ACR57831.1	ATGGACCAACAACAATGGAGTCAGTACTCTAGTTGCTGGCCAGTTT GCGCTCCCATCGCACGCCACGTTTGGCCTGGGAGATCGCGTGCACAAG AAATCTGGCGCCGCTTGGCAGGGTCAAGTTGTCGGGTGGTACTGCAC AAAAGTACCCCTGAAGGCTATGCCGTCGAGTCCGAGTCTACCCCGG TTCAGTACAGATTTATCCTGTGGCTGCGCTTGAACGCGTGGCCTGA	MDQHNNGVSTLVAGQF ALPSHATFGLGDRVRKKS GAAWQQQVVGWYCTKL TPEGYAVESHPGVSQI YPVAALERVA

<i>dfrb4</i>	ALF62656.1 ^a	ATGAATGAAGGAAAAAATGAGGTCAGTACTTCAGCTGCTGGCCGGTTC GCATTCCCATCAAACGCCACGTTTGCCTTGGGGGATCGGTACGCAAG AAGTCTGGCGCTGCTTGGCAGGGGCGCATTGTCCGGTGGTACTGCAC AACACTTACCCCTGAAGGCTACGCCGTCGAGTCCGAATCTCACCCAGG CTCAGTCCAGATTTATCCCATGACTGCGCTTGAACGGGTGGCCTGA	MNEGKNEVSTSAAGRFA FPSNATFALGDRVRKKS AAWQGRIVGWYCTLLTP EGYAVESESHPGSVQIYP MTALERVA
<i>dfrb5</i>	AAX46054.1	ATGGACCAAGGCAGAAAGTGAAGTCAGTAATCCAGTTGCTGGCCAGTTT GCGTCCCTTCAAACGCCCGCTTCCGGAATGGGAGATCGGTGCGCAA GAAATCTGGCGCCGCTTGGCAAGGCCAGATTGTCCGGTGGTACTGCA CAAATTGACCCCTGAAGGCTACGCTGTCGAGTCTGAGGCTCACCTG GCTCGGTACAGATTTATCCTGTTGCGGCCTGGAACGCATCAACTGA	MDQGRSEVSNPVAGQF AFPSNAAFMGDRVRK SGAAWQQQIVGWYCTK LTPEGYAVESEAHPGSVQ IYPVAALERIN
<i>dfrb6</i>	ADO00942.1	ATGGACCAAGGTAGCAATGAAGTCATTAATCCAGTCGCTGGCCAGTTT GCGTCCCATCGAACGCCACGTTTGGTATGGGAGATCGGTGCGCAA GAAATCTGGCGCCGCTTGGCAAGGTCAGATTGTCCGGTGGTACAGCA CAAAGTTGACCCCTGAAGGCTACGCTGTCGAGTCTGAGGCTCACCTG GCTCGGTACAGATTTATCCTGTTGCGGCCTGGAACGCCTCAACTGA	MDQGSNEVINPVAGQF ASPSNATFGMGRVRK SGAAWQQQIVGWYCTK LTPEGYAVESEAHPGSVQ IYPVAALERN
<i>dfrb7</i>	ADB54781.1	ATGGACCAAGGTAGCAATGAAGTCGGTAATCCAGTTGCGGGCCAGTTT TCGTTCCCATCGAACGCCCGCTTGTAGTATGGGAGATCGGTGCGCAAG AAATCGGGCGCCGCTTGGCAAGGTCAGATTGTCCGGTGGTACTGCAC AAAGTTGACCCCTGAAGGCTACGCTGTCGAGTCTGAGGCTCACCTGG CTCGGTACAGATTTATCCTGTTGCGGCCTGGAACGCATCAACGGAGT TCAAGGTTGA	MDQGSNEVGNPVAGQF SFPSNAAFMGRVRK SGAAWQQQIVGWYCTK LTPEGYAVESEAHPGSVQ IYPVAALERINGVQ
<i>dfrb9</i>	AGM20434.1	ATGAATCAAAGTAGCAATTGCATCAGCACTCCAGTTGTTGGACAGTTT GCGCTGCCATTTCAACCCACGTTTGGCCTGGGAGATCGGTACGCAAG AAGTCTGGCGCCGCTTGGCAAGGTAAGTTGTCCGGTGGTACTGCACA AAATTAACCCCTGAAGGCTACGCGGTCGAGTCCGAAGCTCATCCAGGC TCAGTGCAGATTTATCCTGTTGGCTGCGCTTGAACGCCTGGCCTAA	MNQSSNCISTPVVQFA LPFQPTFGLGDRVRKKS AAWQGVVGVWYCTKLT PEGYAVESEAHPGSVQIY PVAALERVA
<i>dfrb10</i>	ALZ46148.1	ATGGATCAAAGTAGCAATGAAGTCAGCACTCCAGTTGCTGGCCAGTTT GCGTCCCATTCGCGCCACGTTTGGCCTGGGAGATCGGTACGCAAG AAATCTGGCGCCGCTTGGCAAGGTCAGTTGTCCGGTGGTACTGCACA AAACTGACCCCTGAAGGCTATGCAAGTCCGAGTCTCACCCAGGC TCAGTACAGATTTATCCTGTTGGCTGCGCTTGAACGCCTGGCCTAA	MDQSSNEVSTPVAGQFA LPLRATFGLGDRVRKKS AAWQGVVGVWYCTKLT PEGYAVESESHPGSVQIY PVAALERVA
<i>dfrb11</i>	PKO69073.1	ATGGATCAAAGTAGTAAAGAGTTGGCACTCCCGTTGTTGGCCAGTTT GCACTCCCGTCGCACGCCACGTTTGGCCTGGGAGACCGCTTCCGCAAG AAATCGGGCGCCGCTTGGCAGGGTCAAGTTGTGGGCTGGTATTGCAC AAAGCTGACCCCTGAAGGCTATGCCGTCGAGTCCGAGTCTCACCCAGG CTCGGTACAAATTTATCCAGTGAATGCGCTTGAACGCCTGGCCTGA	MDQSSKEVGTVPVQFA LPSHATFGLGDRVRKKS AAWQGVVGVWYCTKLT PEGYAVESESHPGSVQIY PVNALERVA

^a This sequence differs from CARD's reference sequence ABY55281.1 at the highlighted region (here Gly instead of Asp). Since we previously characterized DfrB4 using ALF62656.1 sequence¹, we used it in this study.

3.12 Supplementary references

- (1) Toulouse, J. L.; Shi, G.; Lemay-St-Denis, C.; Ebert, M. C. C. J. C.; Deon, D.; Gagnon, M.; Ruediger, E.; Saint-Jacques, K.; Forge, D.; Vanden Eynde, J. J.; Marinier, A.; Ji, X.; Pelletier, J. N. Dual-Target Inhibitors of the Folate Pathway Inhibit Intrinsically Trimethoprim-Resistant DfrB Dihydrofolate Reductases. *ACS Med. Chem. Lett.* 2020, *acsmedchemlett.0c00393*.
<https://doi.org/10.1021/acsmedchemlett.0c00393>.

Chapitre 4. Caractérisation enzymatique et biophysique d'homologues environnementaux des DfrB

Préface

Le chapitre précédent a confirmé la présence en clinique des DfrB dans une variété de contextes génomiques bactériens. L'identification fortuite et la caractérisation expérimentale d'un gène homologue aux DfrB identifié dans un échantillon environnemental dans ce même chapitre nous a incité à entreprendre l'exploration d'homologues plus distants aux DfrB pour étendre nos connaissances sur leur distribution et évolution. En utilisant l'identifiant Pfam des DfrB (correspondant à leur profil de modèles de Markov cachés, voir Chapitre 1), nous avons exploré la base de données protéiques UniProtKB. Les protéines homologues les plus distantes identifiées par cette méthodologie ont été caractérisées expérimentalement pour identifier les similarités entre ces protéines putatives provenant d'échantillons environnementaux et les enzymes DfrB découvertes en clinique.

Cette première étape de caractérisation approfondie d'homologues des DfrB a été essentielle à notre compréhension de leur évolution. Alors que les DfrB avaient été découvertes en clinique suite à l'introduction du triméthopime dans les années 1960s, des protéines homologues venant de contextes environnementaux divers partagent les mêmes caractéristiques enzymatiques et phénotypiques que les DfrB cliniques. Ainsi, cette recherche suggère que la capacité des DfrB à réduire le dihydrofolate n'a pas émergé suite à l'exposition par le triméthopime, mais avant celle-ci.

Ce chapitre correspond à un article invité par le journal *Philosophical Transactions of the Royal Society B* dans le cadre de l'édition spéciale intitulée « Reactivity and Mechanism in Chemical and Synthetic Biology ». Des modifications mineures ont été apportées à la version incluse dans cette thèse.

Prof. Joelle N. Pelletier, Dre Janine N. Copp, Dre Lorea Alejaldre et moi-même avons conçu cette étude. Dre Janine N. Copp a identifié informatiquement les séquences homologues. Dre Donya Valikhani, Katia Hitache (alors stagiaire au B.Sc. sous ma supervision) et moi-même avons réalisé le clonage des constructions. Dre Lorea Alejaldre et moi-même avons entrepris la caractérisation du phénotype des homologues des DfrB. J'ai purifié les protéines, caractérisé l'activité enzymatique des homologues et de leurs variants et réalisé les expériences de chromatographie d'exclusion stérique. J'ai réalisé les expériences de dichroïsme circulaire dans le laboratoire de Prof. Nicolas Doucet, guidée par Myriam Létourneau. Zakaria Jemouai, dans le laboratoire de Prof. Christian Baron, a réalisé les expériences de microscopie

électronique et de SEC-MALLS. Kiana Lafontaine a réalisé les expériences de thermostabilité et fait la détection de l'activité dihydrofolate réductase en lysat. Maxime St-Aubin a réalisé les essais d'inhibition sous la supervision de Kiana Lafontaine. Nuwani W. Weerasinghe, du laboratoire de Prof. Christopher J. Thibodeaux, a fait les expériences de native MS, avec mon aide. J'ai conçu les figures, et j'ai écrit le manuscrit avec Prof. Joelle N. Pelletier. Tous les co-auteurs ont lu et participé à la révision de ce manuscrit.

Article de recherche 2. A conserved SH3-like fold in diverse putative proteins tetramerises into an oxidoreductase providing an antimicrobial resistance phenotype

Claudèle Lemay-St-Denis ^{1,2,3}, Lorea Alejaldre ^{1,2,3}, Zakaria Jemouai ³, Kiana Lafontaine ^{1,2,3}, Maxime St-Aubin ^{1,2,3}, Katia Hitache ^{1,2,3}, Donya Valikhani ^{1,2,4}, Nuwani W Weerasinghe ⁵, Myriam Létourneau ^{1,6}, Christopher J. Thibodeaux ⁵, Nicolas Doucet ^{1,6}, Christian Baron ^{3,7}, Janine N. Copp ⁸ et Joelle N. Pelletier ^{1,2,3,4}

¹ PROTEO, The Québec Network for Research on Protein, Function, Engineering and Applications, Québec, QC, Canada

² CGCC, Center in Green Chemistry and Catalysis, Montréal, QC, Canada

³ Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, QC, Canada

⁴ Chemistry Department, Université de Montréal, Montréal, QC, Canada

⁵ Department of Chemistry and Centre de Recherche en Biologie Structurale, McGill University, Montréal, QC, Canada

⁶ Centre Armand-Frappier Santé Biotechnologie, Institut National de la Recherche Scientifique (INRS), Université du Québec, Laval, QC, Canada

⁷ Department of Microbiology, Infectiology and Immunology, Université de Montréal, Montréal, QC, Canada

⁸ Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada

*Correspondence: joelle.pelletier@umontreal.ca

Philosophical Transactions of the Royal Society B

DOI : [10.1098/rstb.2022.0040](https://doi.org/10.1098/rstb.2022.0040)

© The Royal Society 2023

4.1 Abstract

We present a potential mechanism for emergence of catalytic activity that is essential for survival, from a non-catalytic protein fold. The type B dihydrofolate reductase (DfrB) family of enzymes were first identified in pathogenic bacteria because their dihydrofolate reductase activity is sufficient to provide trimethoprim resistance. DfrB enzymes are described as poorly evolved as a result of their unusual structural and kinetic features. No characterized protein shares sequence homology with DfrB enzymes; how they evolved to emerge in the modern resistome is unknown. In this work, we identify DfrB homologues from a database of putative and uncharacterized proteins. These proteins include a SH3-like fold homologous to the DfrB enzymes, embedded in a variety of additional structural domains. By means of functional, structural and biophysical characterisation, we demonstrate that these distant homologues and their extracted SH3-like fold can display dihydrofolate reductase activity and confer trimethoprim resistance. We provide evidence of tetrameric assembly and catalytic mechanism analogous to that of DfrB enzymes. These results contribute the first insights into a potential evolutionary path taken by this SH3-like fold to emerge in the modern resistome following introduction of trimethoprim.

4.2 Introduction

How does enzyme function arise to ensure host survival when faced with metabolic stress? There are many examples of ancient and modern evolutionary paths that have given rise to resistance mechanisms in response to exposure to xenobiotic compound¹⁻⁵. The best described mechanism for evolution of new enzyme function is the duplication of promiscuous enzymes followed by their divergence to improve an activity required for survival⁶⁻¹⁰. Catalytic specialisation from promiscuous enzymes results from increased affinity, active site rearrangement, allosteric changes and altered protein dynamics¹¹⁻¹⁵. Examples span from the evolution of β -lactamases from DD-peptidases three billion years ago^{1,16,17} to the recent evolution of the AtzA atrazine dechlorinase from TriA melamine deaminase following introduction of the synthetic pesticide atrazine into the environment³.

In contrast, the emergence of catalytic activity that is essential for survival from a non-catalytic protein fold is rare, with few documented examples^{18,19}. Nonetheless, evolution of efficient and stereospecific catalysts from non-catalytic binding proteins has been demonstrated²⁰⁻²². The evolution of a chalcone isomerase from a noncatalytic ancestral protein has been recapitulated, where the successive inclusion of ‘founder’ substitutions at the binding site conferred increasing, initial activity followed by progressive modification of distal residues for fine-tuning purposes²³. Similarly, successive introduction of catalytic residues in the binding site of an ancestral solute-binding protein yielded cyclohexadienyl dehydratase activity that improved upon reshaping by remote substitutions²⁴. In another example, evolution of the organomercurial

lyase activity of MerB occurred through gene duplication and diversification of the TRASH non-enzymatic domain, which has a related metal-binding function ^{25,26}.

Here, we present evidence supporting a further example of evolution of a binding domain into a catalyst. The NADPH-dependent reductase activity of the type B dihydrofolate reductases (DfrB) enables resistance against the antimicrobial trimethoprim (TMP) ²⁷. As opposed to many antimicrobials such as the β -lactams and aminoglycosides, TMP is an entirely synthetic molecule that was clinically introduced in the 1960's and has since been used worldwide in clinical, veterinary and livestock industries settings, including application in preventive measures ²⁸⁻³⁰. As a result, the wide environmental dissemination of TMP has given rise to the rapid appearance of resistance mechanisms ²⁸.

DfrB enzymes are formed by homotetramerisation of a SH3-like fold, consisting of a 60-residue β -barrel. SH3-like folds are highly versatile binding domains that have been shown to interact with molecules ranging from peptides and metals to DNA and RNA ^{31,32}. SH3-like folds have been reported within more than a dozen prokaryotic proteins with diverse binding functions, including protein-protein mediation in the antirepressor CarS ³³ and DNA binding in HIV integrase ³⁴. SH3-like folds have rarely been described to possess catalytic function; to the best of our knowledge, the only examples reported are the type 1 signal peptidases and DfrB enzymes ³².

Evidence supporting the hypothesis that DfrB enzymes may have evolved from a non-catalytic fold is found in their unusual – even qualified as primitive ³⁵ – catalytic mechanism. Whereas the apparent binding constants (K_M) of DfrB enzymes are physiologically relevant, the hydride transfer rate of 1.3 s^{-1} qualifies it as slow among enzymes involved in nucleotide metabolism ^{36,37}. As a result, their catalytic efficiencies are two orders of magnitude lower than the ubiquitous microbial FolA dihydrofolate reductases that are the target of TMP ^{35,38}.

While ten members of the DfrB family have been reported ²⁷, the evolutionary path that has brought DfrB enzymes to the modern resistome is unclear. Aside from sharing the same catalytic function, the homotetrameric, 60-residue β -barrel DfrB enzymes have no structural or evolutionary properties in common with other known dihydrofolate reductases (Dfr), since the ubiquitous FolA and their homologues are monomers of 150 to 190 residues, belonging to the α/β class of proteins (Figure S4.1, Figure S4.2) ^{39,40}.

Here, we identify and characterise DfrB homologues to provide initial insights into the evolution of a SH3-like fold that provides a powerful antimicrobial resistance mechanism. We have identified 30 proteins exhibiting significant sequence homology to the well-characterized DfrB1, by means of searches in a database of predicted and uncharacterized proteins. Characterisation of five putative homologues sharing 10-80% global sequence similarity with DfrB1 revealed that four of these homologues catalysed

dihydrofolate reduction and conferred strong resistance to TMP. Biophysical and kinetic characterisation of the most active distant homologue DfrB-H5 suggest that there is conservation in their ability to multimerize and their catalytic mechanism with the DfrB family. This work unveils a potential mechanism by which a SH3-like fold procures catalytic activity that has become essential for survival in the recent context of exposure to TMP.

4.3 Results

DfrB1, the best characterized member of the 10-member DfrB family (Figure S4.3), is active as an obligate homotetramer, where all four protomers participate in forming the enzyme's central, highly symmetrical active site (Figure 4.1). This voluminous, hourglass-shaped active-site tunnel³⁵ accommodates the substrate dihydrofolate (DHF) and the reducing cofactor NADPH. The active-site residues are known as the conserved VQIY motif³⁸, spanning from V66 to Y69 on the B4 strand. These residues, which are not directly involved in the catalytic event, form a hydrogen bond network both at the floor and ceiling of the active site, forming binding sites that properly orient the reactive groups on DHF and NADPH for the hydride transfer event by a proximity-based catalytic mechanism (Figure S4.4)⁴¹. The putative lack of transition state stabilisation by active site residues is again suggestive of a primitive, or poorly evolved, catalytic mechanism.

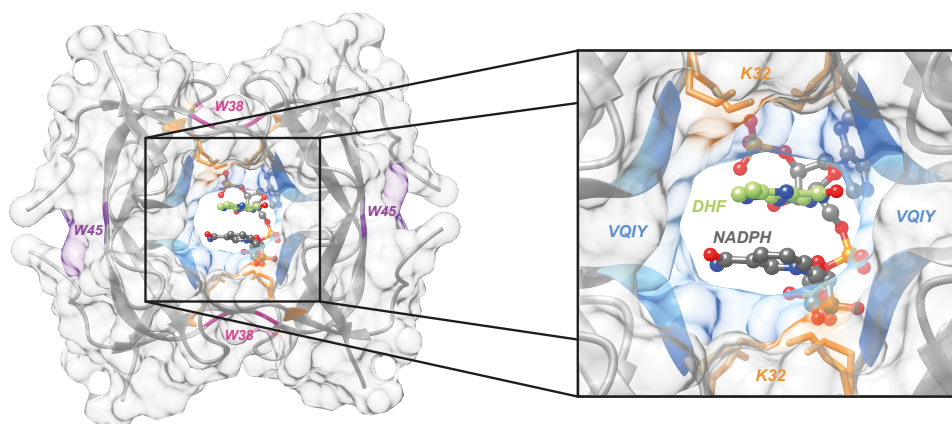


Figure 4.1. Overview of DfrB1 structure and key residues.

The DfrB1 homotetramer (PDB 2RK1) complexed with DHF substrate (carbons in green, only the pterin group is resolved), and NADP⁺ (carbons in grey). The VQIY active site motif is coloured (from dark to light blue) for each protomer. K32 residues that bind the negative charges of DHF and NADPH, are shown as orange sticks (two conformations are shown). Residues W45 at the monomer-monomer interfaces and W38 at the dimer-dimer interfaces, are coloured in purple and pink, respectively.

The symmetrical and tetrameric organisation of the active site make the binding site promiscuous; DHF and NADPH occupy an identical space in the central tunnel. The tunnel has four identical surfaces (Figure 4.1) which together, form a single active site⁴¹. Thus, a single substitution of an active-site residue results

in the simultaneous modification of all four ‘faces’ of the active-site cavity, such that single amino acid substitutions at the active site of DfrB1 are largely deleterious⁴²⁻⁴⁴. This poses a clear disadvantage with respect to natural evolution of a highly adapted catalyst. Nonetheless, because there is no direct catalytic involvement of any residue, fully functional variants of DfrB1 have been engineered where three or all four of the VQIY-motif residues were substituted⁴⁴. That study confirmed that none of the VQIY residues are strictly essential, and that catalysis requires a protein environment conducive for the direct hydride transfer from NADPH to DHF.

4.3.1 Identification of distant homologues of the DfrB enzymes

The protomer unit of DfrB1 is composed of one highly conserved SH3-like fold within the DfrB family, preceded by a poorly conserved, disordered N-terminus (Figure S4.3). This SH3-like fold has no evolutionary homology to any characterized protein. No distant homologues in the UniProtKB/Swiss-Prot database of functionally annotated sequences are identified for DfrB1 according to PSI-BLAST, using an E-value threshold of 10^{-3} ; in contrast, applying the same method to the *E. coli* Fola identifies 90 characterized sequences sharing less than 50% local identity.

Recognizing that standard search tools are inefficient in providing evolutionary insight into the emergence of DfrB enzymes, we searched the complete UniProtKB for DfrB homologues, including uncharacterized proteins. By those means, we identified a total of 68 sequences. They describe 30 non-redundant proteins having sequence similarity with DfrB1 among which 21 are described as putative, as they have only been identified by bioinformatic predictions. Their length ranges from 67 to 463 amino acids. A set of 18 close homologues to the DfrB1 (>80% global sequence similarity) includes nine of the ten known DfrB family members; the remaining 12 homologues are more distantly related (Figure 4.2A, C).

We selected five homologues of DfrB1 for characterisation, which we named DfrB-H2 to DfrB-H6: one close homologue to DfrB1 (DfrB-H2, global similarity of 80%) and four more distantly-related homologues with lower global sequence similarity (between 10 and 31%) due largely to the presence of a diverse array of additional domains (Figure S4.6, Figure S4.7). These DfrB-H proteins are 97 to 463 residues in length (Table S4.1).

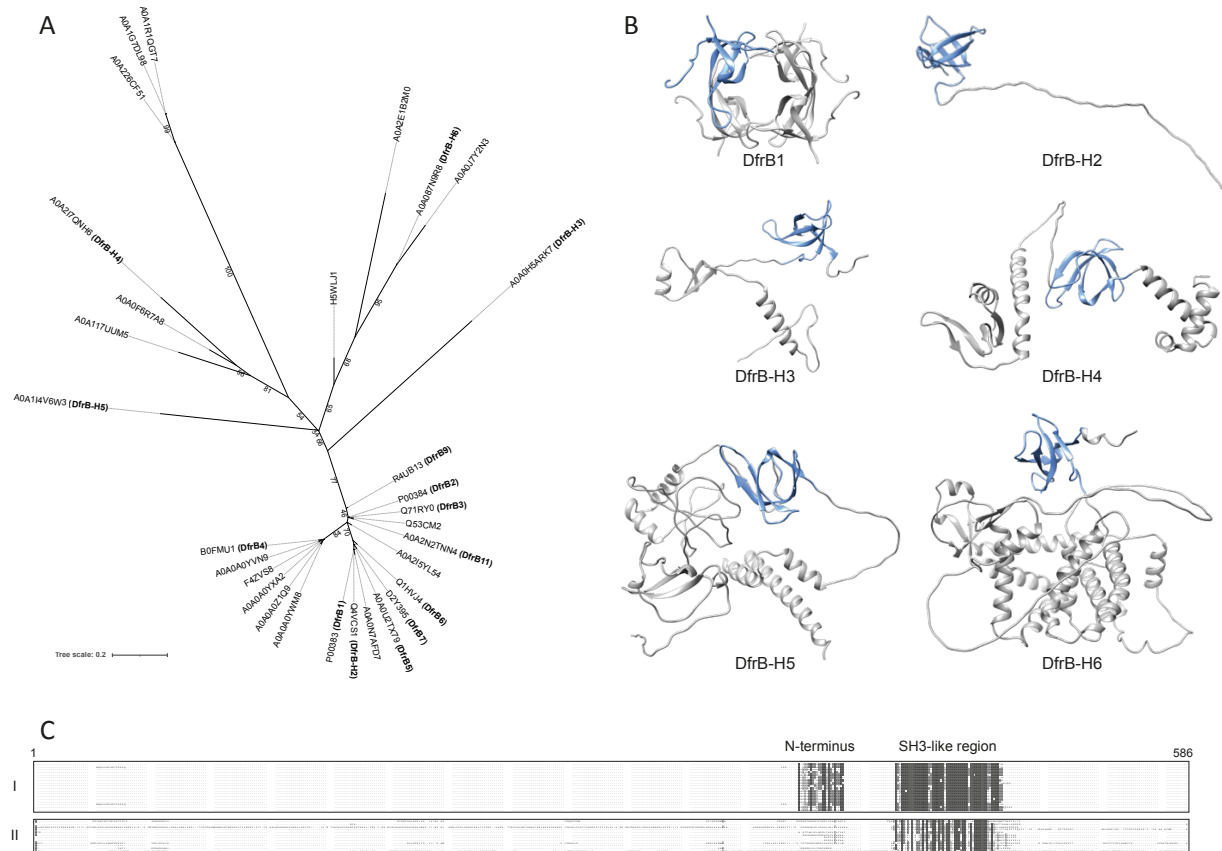


Figure 4.2. Distant homologues of DfrB.

A. Phylogenetic tree (in black) of the 30 non-redundant sequence homologues to DfrB1 in the UniProtKB database was constructed using an alignment of the SH3-like fold (positions 24 to 78 in DfrB1). UniProt IDs are identified for every sequence and connected to their phylogenetic branch with a light grey linker. Bootstrap values are indicated for the branches that separate the main clusters. Nine known DfrB and the five DfrB-H selected for characterisation are annotated in bold. **B.** DfrB1 (PDB 1VIE) homotetramer and DfrB-H structures predicted using ColabFold [45]. The SH3-like folds are colored in blue in all proteins. **C.** Alignment of the 30 sequences. Residues are coloured in grey when identical to the consensus sequence. The cluster I is composed of sequences closely related to DfrB1, while cluster II is composed of distantly related sequences. The alignment was generated by MAFFT [46] and represented by UGENE [47]. The full alignment is presented in Figure S4.5.

DfrB-H2, the closest homologue (Figure 4.2A), has been reported in a number of bacterial contexts. Here, we identified a sequence from *K. pneumoniae* in the genomic context of other genes involved in antibiotic resistance, including the metallo- β -lactamase *blaVIM-1*, the aminoglycoside acetyltransferase *aacA4* and the chloramphenicol acetyltransferase *catB2*. While the 58-residue SH3-like fold is essentially identical to DfrB1 (a single substitution), the 39-residue N-terminus of DfrB-H2 is twice the length and includes the DfrB1 N-terminus. Although the other nine reported DfrB are always identified as 78 residues in length, this gene has two start codons: one for the translation of the 97-residue DfrB-H2, and one for a 78-residue enzyme that is identical to DfrB1 except for one residue at the junction of the N-terminus and the SH3-like

fold. We have no information on the relative expression levels of each gene product *in vivo*. In fact, the well-studied, ‘canonical’ DfrB1 may not be produced in nature in its 78-residue form but may always be accompanied by the longer N-terminus that characterises DfrB-H2. Characterisation of DfrB-H2 will inform on the impact of varying the N-terminus length on DfrB function.

The four more distant homologues, DfrB-H3 to DfrB-H6, are all mainly clustered among genes encoding hypothetical proteins having no predicted function. Among the few genes having a putative function, both DfrB-H3, identified in a *Pseudomonas* phage, and DfrB-H5, identified in *Methylobacterium pseudosasicola*, are found in proximity to a DNA methyltransferase. DfrB-H4, identified in a *Vibrio* phage, is found at a distance of 2 kb from a DNA methylase gene. Finally, DfrB-H6, identified in a *Spingobium* species, is 1.1 kb from a gene for a tyrosine-like recombinase, 1.5 kb from a transcriptional regulator of the LysR family and 2.8 kb from an endonuclease. While the predicted functions of those surrounding genes are related to modification of DNA, the endogenous function of each DfrB-H remains unknown.

Structure prediction for these distant homologues suggests the presence of a SH3-like fold in each of the five homologues (Figure 4.2B). The additional domains in each of the distant homologues share no structural or sequence similarity (Figure S4.8). The SH3-like folds share structural similarity as well as high sequence similarity (67-100%) and identity (55-98%, Figure S4.7), suggesting common ancestry. Notably, the active-site VQIY motif along with the K32 and W38 residues, required for substrate binding and for tetramerisation, respectively (Figure 4.1), are conserved in all DfrB-H (Figure S4.8). However, the W45 monomer-monomer interface residue of DfrB1 (Figure 4.1), that can be substituted with no significant modification of structure or catalytic function⁴⁵, is not conserved.

4.3.2 Investigation of the DfrB-like phenotype in the DfrB-H

For the DfrB-H to display dihydrofolate reductase activity according to the same mechanism as DfrB1, their SH3-like fold must assemble into the homotetramer that characterises DfrB1, thereby forming the active site tunnel. The distant homologues include at least two domains other than the SH3-like fold; as a result, tetrameric assembly might be hindered or precluded. We performed minimal inhibitory concentration (MIC) assays to investigate whether the DfrB-H proteins confer a TMP resistance phenotype consistent with dihydrofolate reductase activity when overexpressed in *E. coli*. We note that DfrB1 and the DfrB-H assayed in this study carried an N-terminal His₆ tag. Surprisingly, DfrB-H2, DfrB-H4, DfrB-H5 and DfrB-H6 provided TMP resistance up to the highest concentration of TMP soluble in 5% methanol (600 µg/mL) (Figure 4.3A, Table S4.2).

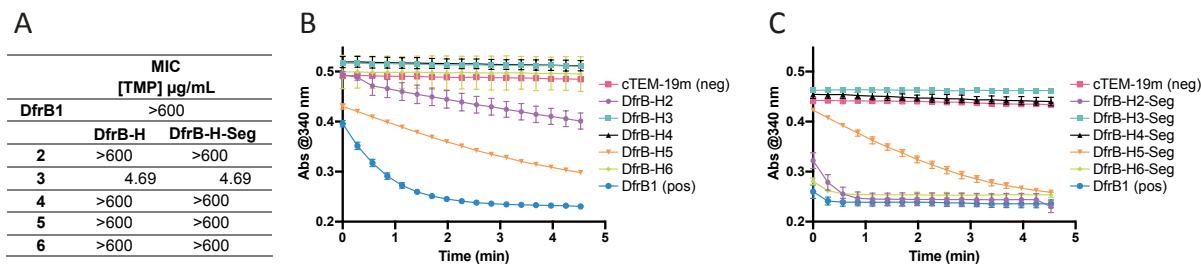


Figure 4.3. DfrB-H proteins display a DfrB1-like phenotype.

A. MIC assays performed in *E. coli* on LB-agar IPTG induction media in triplicate. TMP concentration ranged between 4.7 and 600 $\mu\text{g/mL}$, the highest soluble concentration. **B-C.** Dihydrofolate reductase activity in crude *E. coli* lysate following overexpression of the **B.** DfrB-H and the **C.** DfrB-H-Seg, segments encoding the predicted SH3-like fold of the DfrB-H. A decrease in absorbance at 340 nm indicates depletion of the substrate DHF and the cofactor NADPH. The fastest reaction rates are seen where substrate depletion has occurred even at the initial time point. An absorbance that remains high and constant is indicative of non-detectable dihydrofolate reductase activity. Assays were performed in triplicate and error bars represent standard deviation.

We then conducted assays for dihydrofolate reductase activity using cell lysate from the recombinant *E. coli* strains expressing the DfrB-H. Significant activity was detected only in the lysate of cells expressing DfrB-H2 and DfrB-H5 (Figure 4.3B). Therefore, DfrB-H2 and DfrB-H5 showed both dihydrofolate reductase activity in bacterial lysate and TMP resistance *in vivo*. In contrast, the TMP resistance observed in DfrB-H4 and DfrB-H6 using the more sensitive MIC assay is consistent with dihydrofolate reductase activity too low to be detected in the cell lysates. Consistent with this hypothesis, the DfrB-H2 and DfrB-H5 proteins are readily observed upon overexpression, lysis and resolution by tricine-SDS-PAGE, whereas the DfrB-H3, DfrB-H4 and DfrB-H6 proteins did not express at levels high enough to be visualised (Figure S4.10).

4.3.3 Extracting the homologous SH3-like segments from the DfrB-H

To better evaluate the similarities between the SH3-like fold of the DfrB-H proteins and DfrB1, we generated 78-residue segments named DfrB-H-Seg. The DfrB-H-Seg proteins are composed of the predicted SH3-like fold of the DfrB-H, preceded by the 20-residue N-terminus of DfrB1 (Figure S4.9). In the case of DfrB1, this N-terminus was shown to be essential for expression of the well-folded active protein, although its subsequent removal does not abrogate activity³⁶. Indeed, all DfrB-H-Seg (except for DfrB-H3) could be observed on tricine-SDS-PAGE upon overexpression (Figure S4.10). Notably, removal of the additional domains resulted in visible protein expression of DfrB-H4-Seg and DfrB-H6-Seg, whereas the full DfrB-H4 and DfrB-H6 proteins were not detectable.

MIC assays in *E. coli* overexpressing these DfrB-H-Seg proteins revealed TMP resistance phenotypes similar to their respective full-length protein (Figure 4.2A, Table S4.2), indicating that the resistance phenotype is independent of the additional domains in the DfrB-H proteins. This was also the case for DfrB-H3-Seg protein expression which failed to provide TMP resistance, consistent with the lack of TMP resistance from the full-length DfrB-H3. This suggests that the lack of activity of DfrB-H3 does not result from its additional domains preventing homotetramerisation, but rather from a lack of soluble expression.

The dihydrofolate reductase activity of DfrB-H6-Seg in cell lysate was as high as that of DfrB1, whereas DfrB-H5-Seg displayed lower activity (Figure 4.3C). The differences in activity between the full-length proteins and their extracted domains may result from differing expression levels; in particular, DfrB-H6 lacked visible expression and activity whereas DfrB-H6-Seg appeared as a clear band and showed high activity (Figure S4.10). These results demonstrate that the SH3-like fold of the DfrB-H proteins can suffice to procure TMP-resistant dihydrofolate reductase activity.

4.3.4 Kinetics and inhibition of the DfrB-H5 distant homologue

Having determined that the 41.6 kDa DfrB-H5 and that the independently expressed 10.9 kDa SH3-like fold DfrB-H5-Seg display similar activity in lysates, we compared their kinetic parameters (k_{cat} , K_M^{DHF} , K_M^{NADPH}) to the 11.0 kDa DfrB1 reference protein. DfrB-H5 and DfrB-H5-Seg exhibit similar kinetic parameters as DfrB1 (Table 4.1), demonstrating that the dihydrofolate reductase activity of DfrB-H5 is entirely defined by its SH3-like fold. The Y69L substitution has been shown to reduce ligand binding and decrease the rate of catalysis in DfrB1⁴³. To further investigate the mechanistic similarities between DfrB1 and DfrB-H5, their VQIY active site motives was modified to VQIL. This produced a loss of trimethoprim resistance and a significant reduction in dihydrofolate reductase activity in both DfrB1 and DfrB-H5 (Table 4.1, Table S4.3). The analogous modulation of the DfrB1 and DfrB-H5 catalytic activities by this active site substitution further supports mechanistic similarities.

Table 4.1. Kinetic parameters for the dihydrofolate reductase activity

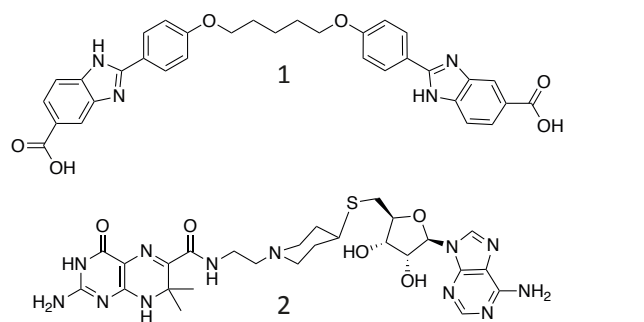
	K_M^{DHF} (μM)	K_M^{NADPH} (μM)	k_{cat} (s^{-1})	k_{cat}/K_M^{DHF} ($\text{s}^{-1} \mu\text{M}^{-1}$)
DfrB1 ^a	8.2 ± 0.11	1.6 ± 0.02	0.83 ± 0.01	0.10
DfrB-H2	8.7 ± 0.6	7.8 ± 0.6	1.35 ± 0.02	0.15
DfrB-H5	21 ± 7	12 ± 1	1.1 ± 0.2	0.05
DfrB-H5-Seg	30 ± 20	20 ± 4	3 ± 1	0.09
DfrB1 Y69L	400 ± 400	170 ± 20	0.03 ± 0.03	8.0 × 10 ⁻⁵
DfrB-H5 Y267L		NA ^b		

^a Reference [44]

^b Trace activity detected

To further probe the structural similarities of the active sites of DfrB1 and DfrB-H5, we characterized the inhibition of dihydrofolate reductase activity of DfrB-H5 by representatives of two distinct classes of DfrB inhibitors. Inhibitor 1 belongs to a class of symmetrical bis-benzimidazoles that has been demonstrated to provide strong inhibition (K_i in the range of 2-62 μM) of DfrB enzymes by binding inside the active-site tunnel^{38,46,47}. The second class of inhibitors (such as inhibitor 2) is composed of bisubstrate composite molecules formed from the DHF and NADPH substructures, and also effectively inhibit the DfrB enzymes (K_i 12-130 μM)³⁸. Here, we show that DfrB-H5 and DfrB1 display essentially indistinguishable affinities for each inhibitor (Table 4.2). Considering that the two classes of inhibitors are structurally unrelated, this finding is consistent with the binding region of DfrB-H5 being structurally analogous to that of DfrB1.

Table 4.2. Inhibition of DfrB-H5 with structurally distinct, DfrB-specific inhibitors. Inhibitor 1 is a bis-benzimidazole molecule, and inhibitor 2 is a bisubstrate molecule.



	1		2	
	K_i (μM)	IC50 (μM)	K_i (μM)	IC50 (μM)
DfrB1	2.0 ± 0.3^a	64 ± 11^a	20 ± 3^b	650 ± 87^b
DfrB-H5	12 ± 6	60 ± 30	30 ± 10	170 ± 60

^a Ref [49]

^b Ref [38]

4.3.5 The DfrB-like domain promotes tetramerization

The functional evidence above supports the hypothesis that the distant homologue DfrB-H5 forms an active site that is structurally analogous to that of DfrB1. We next investigated whether DfrB-H5 assembles into a homotetramer, which is known to be essential for the formation of functional DfrB1. We observe tetrameric arrangements of similar symmetry for both enzymes in negative-stain Electron Microscopy (EM), as well as a similar diameter for the single central tunnel (Figure 4.4A). As expected, the DfrB-H5 tetramer (41.6 kDa per monomer) is 100 Å in diameter, which is significantly larger than the DfrB1 tetramer (58-70 Å; 11.0 kDa per monomer). This data supports the formation of a symmetrical pore that is analogous to the functional DfrB-like active site.

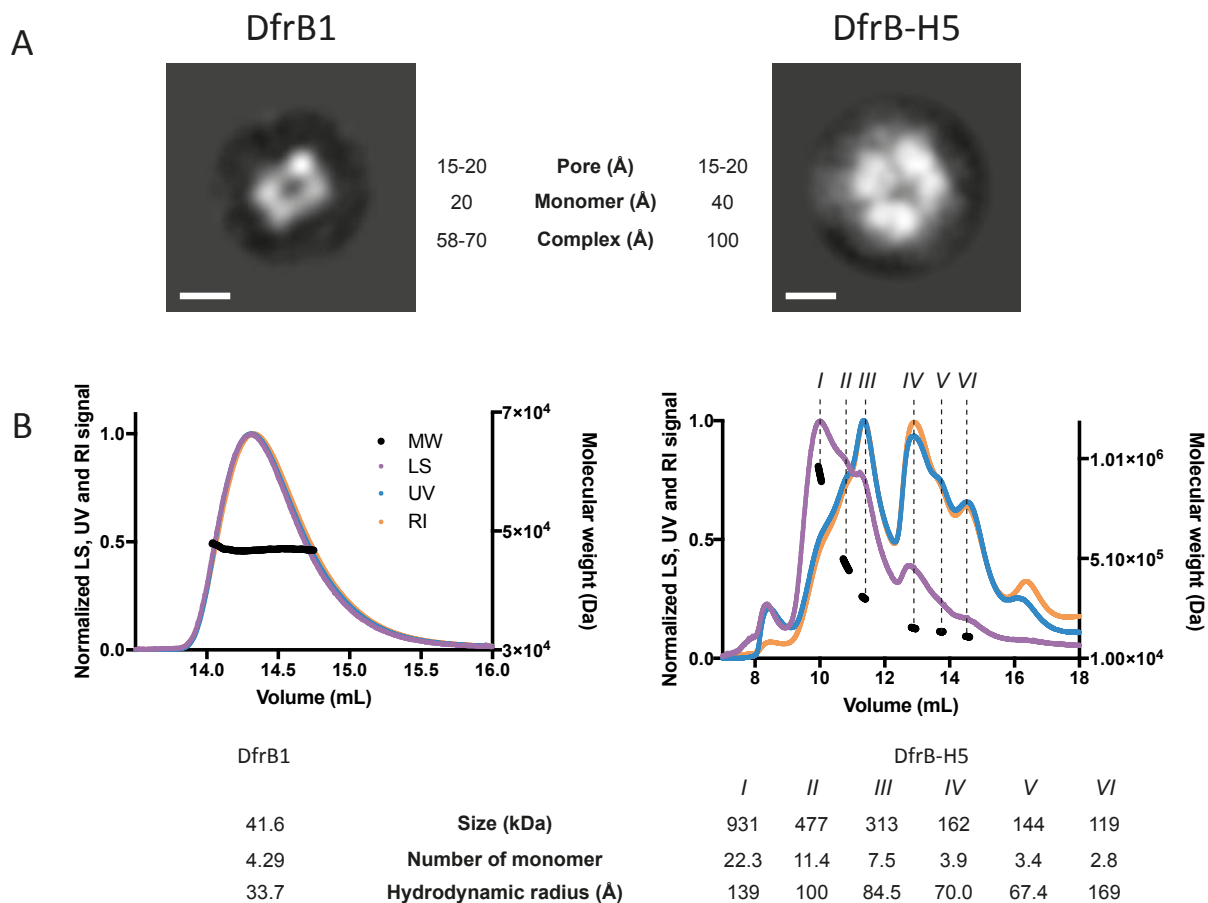


Figure 4.4. Oligomerisation of DfrB-H5 analysed with 2D classification and light scattering.

A. Negative stain electron microscopy images for DfrB1 (left) and DfrB-H5 (right) reveals homotetramerisation of the proteins with pores of similar size. The images for DfrB1 and DfrB-H5 are a superposition of 1013 and 277 particles respectively. The size of each monomer, overall complex and central pore is shown. The white bar corresponds to 50 Å. **B.** Elution profile, with normalized light scattering data (LS) in purple, UV absorbance in blue and differential reflective index data (RI) in orange, of 3.27 mg/mL DfrB1 (left) and 0.57 mg/mL DfrB-H5 (right) shown with the molecular weight estimated by MALLS in black. The size, the estimated number of monomer and the hydrodynamic radius is presented for each peak.

DfrB1 assembled uniformly into apparent tetramers as evidenced by EM (Figure 4.4). In contrast, 2D class averages for DfrB-H5 were consistent with the formation of the tetrameric form (Figure 4.4) as well as of dimers and trimers (Figure S4.11). We further investigated this multimeric assembly in solution, using analytical size exclusion chromatography-multi-angle laser light scattering (SEC-MALLS)⁴⁸ (Figure 4.4B). Once again, DfrB1 formed highly homogenous tetramers. In contrast, multiple heterogenous peaks were detected in the case of DfrB-H5, with the main species being the tetrameric (3.9 monomers) and the octameric (7.5 monomers) forms. We also observed oligomeric species predicted to belong to dodecameric

(11.4) and 24-mer (22.3) particles. Based on the higher order oligomers being multiples of 4 (8-, 12- and 24-mer), we hypothesise that these are complexes of tetramers. Other regions of the protein may contribute to formation of higher order oligomers that are observed. Taken together, results from EM and SEC-MALLS highlight the consistent assembly of DfrB1 into a functional homotetrameric form, whereas DfrB-H5 assembles into various oligomeric arrangements, consistent with the association of multiple tetramers.

In order to form its functional homotetrameric form, two unfolded DfrB1 monomers (M) dimerise; two dimers (D) then dimerise to form a tetramer (T), according to $4M \rightleftharpoons 2D \rightleftharpoons T$ ^{36,45,49}. Each protomer contains a W38 at the dimer-dimer interface (Figure 4.1)⁴⁵. Each dimer-dimer interface also includes two H62 residues; as a result, DfrB1 is mainly in a dimeric state at pH 5 due to protonation of H62, and is predominantly tetrameric at pH 8^{49,50}. The W38F substitution in DfrB1 impedes the formation of functional tetramer, resulting in 10- and 40-fold weaker effective binding to DHF and NADPH, respectively, and a 100-fold reduced catalytic turnover⁴⁵. Nonetheless, the residual activity of W38F DfrB1 suggests a low level of tetramer formation.

Since W38 is conserved throughout the DfrB family and in all DfrB-H (Figure S4.8), the W38F substitution could serve to investigate similarities in the association mechanism of the SH3-like fold in these proteins. We began by further characterising the phenotype and biophysical properties of W38F-substituted DfrB1. First, we confirmed by native mass spectroscopy (MS) that the W38F substitution significantly shifts the oligomeric populations of DfrB1, leaving a scarcely detectable population of tetramers (Figure S4.12). This trace of tetramers apparently accounts for the trace of dihydrofolate reductase activity that is required to confer TMP resistance when overexpressed in *E. coli* (Table S4.3).

To determine whether assembly of the protomers into a tunnel-forming tetramer is analogous in DfrB1 and DfrB-H5, we substituted the conserved tryptophan of DfrB-H5 corresponding to W38 in DfrB1 (W236). We attempted to perform native MS measurements on DfrB-H5 and its resulting W236F variant; it was not possible to obtain signals, presumably due to the poor ionisation of the larger DfrB-H5 protein. Although the W236F substitution did not yield significant changes in the band pattern of DfrB-H5 on native PAGE (Figure S4.13), it gave rise to a trimethoprim sensitive phenotype (Table S4.3), indicative of a greater loss of dihydrofolate reductase activity than W38F DfrB1. This supports the hypothesis that the SH3-like fold of DfrB-H5 directs protein-protein interactions between protomers to form a tetramer analogous to that of DfrB1.

SEC analysis showed a clear shift of the main peak for DfrB-H5, centered at 1.31 mL, to 1.52 mL for W236F DfrB-H5, demonstrating a change in the predominant states of multimerisation (Figure S4.14). In

both cases, shoulders are observed on either side of the main peak. The poor separation of the species is consistent with conformational exchange within this population occurring on the timescale of the SEC experiment. Overall, this demonstrates that the conserved tryptophan plays a critical role in ensuring sufficient dihydrofolate reduction when DfrB-H5 is expressed in bacteria. The impact of the conserved W→F mutation is not identical in DfrB1 and DfrB-H5, which may indicate that other regions of DfrB-H5 participate in multimerisation. Validation of this claim must await further studies.

Cumulatively, this biophysical data is compatible with a conserved role for the SH3-like fold in DfrB1 and in its distant homologue DfrB-H5. This region, characterized by high sequence conservation (69% similarity between DfrB1 and DfrB-H5, Figure S4.7), appears to be sufficient to drive tetramerisation even in the context of the much larger DfrB-H5 protein. The 58-residue SH3-like fold of DfrB-H5 is embedded within a larger protein architecture (residues 219 to 276 of a 365-residue protein), flanked by predicted α and β domains (Figure 4.2B). The SEC-MALLS and the native PAGE data agree on the formation of defined, higher order oligomers. SEC-MALLS identifies the dominant forms of DfrB-H5 as being tetrameric and octameric, with dodecamer readily identified and a trace of 24-mer (Figure 4.4B). This suggests the formation of tetramers, as well as dimers and trimers of tetramers (octamers and dodecamers), along with a trace of dimers of dodecamers. This is consistent with the SH3-like fold being the dominant force in multimer assembly, with other interactions of weaker strength potentially mediated by the other domains in DfrB-H5 not shared with DfrB1.

4.3.6 Biophysical properties of the SH3-like fold are similar in DfrB1 and DfrB-H5-Seg

Using circular dichroism (CD), we verified whether the extracted SH3-like segment of the distant homologues display similar properties. The His₆-tagged DfrB1 includes 33% of structured regions and 67% of loops and the disordered N-terminus. Consistent with this, DfrB1 shows two characteristic shoulders at 210 and 240 nm with a broad minimum between 215 and 225 nm at 20°C (Figure 4.5A). The minimum at 203 nm is characteristic of the unstructured regions of DfrB1. Heating to 95°C induced little change other than a loss in definition (Figure 4.5A). Considering the many factors that can cause a signal change upon heating an oligomeric enzyme, we hypothesised that this 215-230 nm signal is a signature of homotetramerisation. Indeed, the CD spectrum at 20°C of DfrB1 W38F and DfrB1 at pH 5, both impeded in tetramer formation⁴⁵, did not exhibit this signal (Figure 4.5B).

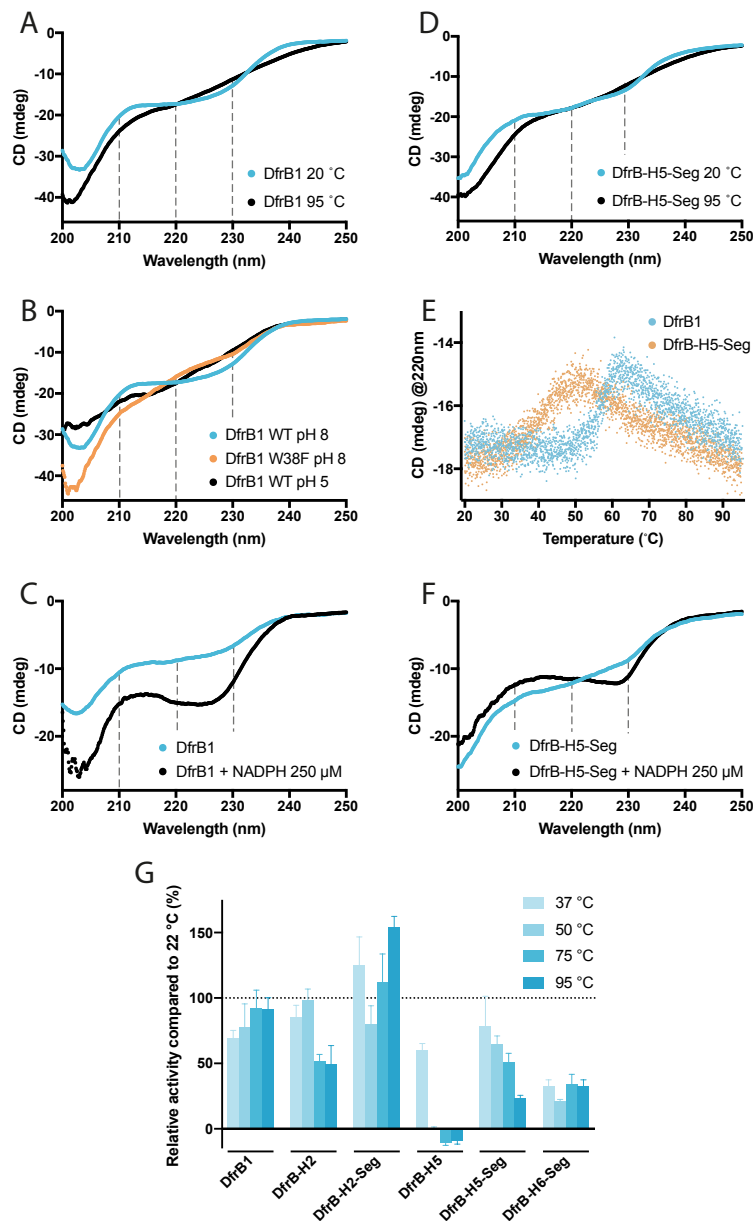


Figure 4.5. Influence of heating on multimerisation and activity.

A. The 215-230 nm signature CD signal of DfrB1 at 20 °C is lost when heated at 95 °C. **B.** The 215-230 nm signal is present only for DfrB1 at pH 8, when homotetramerisation is possible. The W38F substitution in DfrB1 and pH 5 disrupt tetramerisation. CD spectra at 20 °C of DfrB1 (40 μM pH 8), DfrB1 W38F (40 μM pH 8) and DfrB1 (35 μM pH 5). **C.** The 215-230 nm signal is stabilised when DfrB1 is incubated with NADPH. CD spectra of DfrB1 (25 μM) with and without incubation of NADPH 250 μM. **D.** The CD spectra of DfrB-H5-Seg also displays the 215-230 nm signature signal, comparable to DfrB1. The spectra at 95 °C does not present the signature signal. **E.** A signal change at 220 nm is observed at a lower temperature for DfrB-H5-Seg than DfrB1. A minor signal change occurs at 56.9 ± 0.2 °C and 43 ± 1 °C for DfrB1 and DfrB-H5-Seg respectively. **F.** CD spectra of DfrB-H5-Seg (25 μM) with and without incubation with 250 μM NADPH. The 215-230 nm signal is stabilised when DfrB-H5-Seg is incubated with NADPH. **G.** Thermotolerance of dihydrofolate reductase activity for DfrB-H and DfrB-H-Seg proteins with detectable dihydrofolate reductase activity in cell lysate.

The SH3-like fold of DfrB-H5-Seg presented a similar spectral signature between 210-240 nm at 20°C and a similar loss of definition upon heating (Figure 4.5D). This further supports similarity in the mechanism of tetramerisation of the SH3-like fold in DfrB1 and DfrB-H5. Nonetheless, the T_m (220 nm) shifted from 56.9°C in DfrB1 to 43 °C in DfrB-H5-Seg, demonstrating that its homotetramer is less thermostable than that of DfrB1 (Figure 4.5E).

Ligand binding often stabilises protein structure^{51,52}. We investigated whether incubation of DfrB1 with NADPH and folic acid (the air-stable analogue of DHF), resulted in a modification of the CD signal, which would indicate a change in the equilibrium of the oligomeric species. In DfrB1, NADPH and DHF binding are mediated by residues from more than one protomer within the tetrameric tunnel, as observed by crystallography⁴¹ (Figure 4.1). Where the weak binding of folic acid (K_D 120 μ M⁵³) with DfrB1 did not modify the CD signal (data not shown), the strong binding of NADPH (K_D 2.5 μ M⁵³) sharpened the 210 nm shoulder and accentuated the local minimum at 225 nm (Figure 4.5C). The DfrB-H5-Seg protein displayed a parallel, albeit weaker pattern upon NADPH binding (Figure 4.5F). This is consistent with a similar NADPH binding mechanism in DfrB1 and DfrB-H5.

DfrB1 is known to tolerate incubation at high temperature^{38,54}. Other members of the DfrB family display similarly high recovery of activity following heating to 95°C and cooling (Figure S4.15). In fact, DfrB enzymes can be purified from lysate using incubation at 70°C³⁸. We investigated whether the DfrB-H and their extracted segments producing detectable dihydrofolate reductase activity (Figure 4.3B,C) exhibited similar thermotolerance. DfrB1 tolerated heating to 95°C for 10 minutes, recovering 91% activity upon cooling (Figure 4.5G). DfrB-H2 displayed lower thermostability (50% activity recovered after heating to 75°C and 95°C) as a result to the 19 residues added to its 20-residue N-terminus. This is supported by the observation that the extracted segment of DfrB-H2, DfrB-H2-Seg, is thermostable, consistent with it differing from DfrB1 by a single residue. In contrast, DfrB-H5 lost activity in lysate following heating to 50°C or more. This can be attributed to the additional domains of the 365-residue DfrB-H5. Nonetheless, the extracted SH3-like fold of DfrB-H5, DfrB-H5-Seg, displayed considerable tolerance to heating, as did DfrB-H6-Seg. They respectively recovered 23% and 33% of their initial activity following heating to 95°C (Figure 4.5G). The SH3-like fold of each of these proteins thus displays a pattern of heat tolerance reminiscent of that of DfrB1 yet with distinctive features that reflect their sequence differences (Figure S4.8).

4.4 Discussion

The TMP-resistant DfrB proteins were first detected on plasmids of pathogenic bacteria in the 1970's⁵⁵⁻⁵⁷. They have since been described as primitive enzymes, as various lines of evidence demonstrate that their catalytic mechanism had not been optimized by evolution³⁵. A crucial indicator of this primitive mechanism

is the absence of a catalytic acid. As a result, the catalytic mechanism relies on protonation of the DHF-N5 ($pK_a = 2.59$ ⁵⁸) by the solvent (Figure S4.4). This demonstrates a lack of evolutionary fine-tuning, in contrast to the catalytic mechanism of the ubiquitous Fola dihydrofolate reductase family, where a conserved aspartate or glutamate increases the pK_a of DHF-N5, facilitating the hydride transfer from NADPH to the imine⁵⁹.

Databases of characterized proteins yielded no protein homologous to the DfrB family using standard search tools. This begs the question: how has the DfrB family evolved to emerge in the modern resistome? We turned to the investigation of putative proteins to establish a potential evolutionary link with the DfrB family. Our search in the UniProtKB database, which includes putative proteins, yielded only 30 non-redundant proteins with sequence similarity to DfrB1, among which 12 were distant homologues. The predicted SH3-like fold of DfrB-H2 to DfrB-H6 share between 67 and 100% sequence similarity with DfrB1, suggesting common ancestry. Not only is high sequence similarity shared between the β -strands of the SH3-like fold, but both the length and sequence are conserved in the inter-strand loops. This is notable, as inter-strand loops of SH3-like folds tend to differ greatly³².

Remarkably, four of the five DfrB-H, sharing between 10 and 80 % global sequence similarity with DfrB1, displayed the same high TMP resistance phenotype. The DfrB-H5 homologue (sharing 14 % global sequence similarity and 63 % local sequence similarity with DfrB1) displayed clear catalytic activity as well as numerous structural and functional similarities with DfrB1. These include similar pore size resulting from homotetramerisation, the importance of the conserved VQIY motif and K32 to bind the negatively charged groups of DHF and NADPH, and the similar inhibition of both enzymes by two structurally distinct molecules. The primitive catalytic mechanism is likely proximity-based, with the active site orienting the reactive groups of DHF and NADPH for the hydride transfer event⁴¹.

We have thus demonstrated that the DfrB-H proteins, consisting of a predicted SH3-like fold homologous to DfrB enzymes and a variety of additional structural domains, can provide the same antimicrobial resistance phenotype as the DfrB family. Their core architecture is compatible with dihydrofolate reduction by a DfrB-like mechanism. This suggests that the ancestors of the DfrB family could have displayed adventitious dihydrofolate reduction activity embedded in the context of varied and complex protein architectures of yet unknown function. The selective pressure recently provided by TMP could have promoted the extraction of the DfrB-like domain and its integration into the resistome by means of mobile genetic elements. This hypothesis is supported by the extraction of the SH3-like fold of the DfrB-H, where DfrB-H4-Seg, DfrB-H5-Seg and DfrB-H6-Seg display the same phenotype as the DfrB1. This demonstrates that the SH3-like fold of distant DfrB homologues can suffice to provide the phenotype required for survival when challenged with trimethoprim. This is also consistent with the N-terminal extension of DfrB-H2

relative to the nearly identical though shorter DfrB1. In this case, it suggests that recent duplication and diversification of DfrB-H2 could have led to the second Met acting as the only start codon, yielding the 78-residue products that define the modern DfrB family.

Despite having uncovered structural and functional links between the DfrB and the diverse homologues, we have not accrued sufficient information to gain clear insight into the evolutionary origin of the DfrB family. The DfrB-H characterized here have no known native function nor evolutionary background, as searches for their homologues in UniProtKB/Swiss-Prot using PSI-BLAST yielded no significant hit aside from known DfrB members. Although none of the distant DfrB-H belong to the same structural or evolutionary family, their high sequence homology within the SH3-like fold supports a relation resulting from divergent evolutionary relationship (rather than convergent evolution) for the DfrB-like domain. We envision that tapping into the information captured in metagenomic databases will be essential to recapitulate the evolutionary path of the DfrB family towards the modern resistome.

4.5 Methods

4.5.1 Identification of the homologues

The homologous sequences were gathered from UniProtKB by searching for the Pfam family designation 06442 [UniProt release 2018_06]^{60,61}. The resulting dataset of 68 sequences (30 non-redundant sequences) ranged from 67 to 463 amino acids in length. Non-redundant sequences were aligned with MAFFT⁶². The phylogenetic tree was generated using the alignment of the SH3-like sequence (positions 24 to 78 in DfrB1) using IQ-TREE (ultrafast bootstrap analysis, 1000 alignments)⁶³. The tree was represented using iTOL⁶⁴. Structure prediction was performed by ColabFold, using the relaxed option⁶⁵.

Five sequences, named DfrB-H, were selected from multiple sequence alignment and phylogenetic reconstructions, to ensure broad sampling of the Pfam06442 sequence space^{66,67}. These sequences were codon optimised for *E. coli* expression and synthesised for cloning purposes.

4.5.2 Cloning

The DfrB-H, all with a N-terminal His₆-tag, were ordered via TwistBioscience in pET28a vectors. Subcloning of cTEM-19m was performed as previously described⁶⁸. His-tagged DfrB3, DfrB5 and DfrB7 were subcloned in pET24 as previously reported³⁸, and His₆-DfrB4 was obtained as reported previously⁴⁷. DfrB1 with and without His₆ in pET24 was obtained as previously reported^{38,47}. Otherwise indicated, all mentions of DfrB1 refers to His₆-DfrB1. Mutations W38F and Y267L were respectively introduced into DfrB1 and DfrB-H5 with Phusion Plus polymerase (Thermo) in a two-step reaction (1 min at 98°C, 30 cycles of 15 s at 98°C and either 2 or 3 min at 72°C, 5 min at 72°C). The primers used were as follows:
DfrB1-W38F-F (5'-CGCCGCCTTCCAAGGTCAGATTG-3'), DfrB1-W38F-R (5'-

CCGGATTTCTTGCGCACGCG-3'), DfrB-H5-Y267L-F (5'-TGTCCAAATTTTGCCGATCGCAGC-3'), DfrB-H5-Y267L-R (5'-CTACCAGGTTTCACGTTCTGACTCG-3'). Templates were digested using DpnI (NEB) O/N at 37°C. Mutations Y69L and W236F were respectively introduced in DfrB1 and DfrB-H5 using the QuickChange Lightning kit, with either 3 min or 3 min 40 s elongation time. The primer used were designed according to the kit: DfrB1-Y69L-F (5'-GCTCAGTACAGATTTTACCTGTTGCGGCGCTTGAACGCA-3'), DfrB1-Y69L-R (5'-GCGCCGCAACAGGTAAAATCTGTACTGAGCCTGGG-3'), DfrB-H5-W236F-F (5'-CGCAAACCTAAAGGTTCTAGTTTCCAGGGAGTAGTGG-3'), DfrB-H5-W236F-R (5'-CTACTCCCTGGAACTAGAACCTTTAGTTTTGCGCAC-3'). Reactions were transformed into chemically competent *E. coli* DH5 α .

DfrB-H2-Seg, DfrB-H3-Seg and DfrB-H6-Seg were generated using a modified version of restriction-free (RF) cloning⁶⁹. First, the pET24 backbone with the first 20 amino acids of the DfrB1 gene was amplified using 100 ng of template and following a 3-step protocol (1 min at 98°C, 18 cycles of 30 s at 98°C, 50 s at 5°C below melting temperature and 1min/kb at 68°C, and 10 min at 68°C) using the PfuUltra polymerase (Agilent). The following primers were used: pET24-F (5'-TAAAAGCTTGCGGCCGCACTC-3'), Nterm-R (5'-CGATGGGAATACAAAATTGCCAGCAAC-3'). Then, the megaprimer was generated by amplifying the segments of DfrB-H3 and DfrB-H6 with a 3-step protocol (30 s at 98°C, 30 cycles of 30 s at 98°C, 30 s at 55°C and 20 s at 72°C, and 10 min at 72°C) using the Phusion polymerase (Thermo). The primers used are as follows: DfrB-H2-Seg-F (5'-CCAAGACTACAAAGACGATGACGACAAGATGGAACGTTCTAGCAATGAGG-3'), DfrB-H2-Seg-R (5'-GTGCGGCCGCAAGCTTTTAATTTATGCGTTCCAAGGCTGC-3'), DfrB-H3-Seg-F (5'-GTTGCTGGCAATTTTGTATTCCCATCGCAGGGAAAATTCCGCATGG-3'), DfrB-H3-Seg-R (5'-CGAGTGCGGCCGCAAGCTTTTACATCCACTGACGCCATCT-3'), DfrB-H6-Seg-F (5'-GTTGCTGGCAATTTTGTATTCCCATCGGTGGGCAAATTCAGCGAG-3'), DfrB-H6-Seg-R (5'-CGAGTGCGGCCGCAAGCTTTTAATGTGAAAGCCGAAGGGC-3'). Templates were digested by DpnI (NEB) for 2 h at 37°C and reactions were cleaned using the Monarch PCR & DNA Cleanup Kit (NEB). With the cleaned megaprimers and the pET24 backbone, the DfrB-H-Seg were assembled using the 2-step Secondary PCR protocol from rf-cloning.org⁷⁰. Reactions were cleaned using the Monarch PCR & DNA Cleanup Kit (NEB) and transformed into chemically competent *E. coli* DH5 α . DfrB-H4-Seg and DfrB-H5-Seg were ordered from TwistBioscience in pET24. All sequences were confirmed by DNA Sanger sequencing (Genome Quebec platform at Sainte-Justine Hospital).

4.5.3 Minimal inhibitory concentration

MICs were determined in triplicates according to Wiegand et al.⁷¹ using both the agar and broth microdilution method. Briefly, *E. coli* BL21(DE3) cells expressing DfrB-H, DfrB-H-Seg, positive control DfrB1 and negative control cTEM-19m⁶⁸ were propagated overnight in Luria-Bertani (LB) medium with 50 µg/mL kanamycin. For the agar method, an inoculum of 10⁴ colony forming units (cfu) was spotted on LB agar plates with 0.25 mM IPTG (ThermoFisher) and TMP (Sigma) in 2-fold concentration steps up to 600 µg/mL; the latter is the highest concentration of TMP soluble in a final concentration of 5% methanol. The TMP concentration inhibiting bacterial growth following overnight incubation at 37°C was considered to be the minimal inhibitory concentration. For the broth method, in 96-well plates, LB media was inoculated with 10⁵ cfu/mL, with 0.1 mM IPTG and TMP. Minimal inhibitory concentrations were determined as described above.

4.5.4 Dihydrofolate reductase activity in lysate

The various DfrB proteins (DfrB1, DfrB-H, DfrB-H-Seg) and negative control cTEM-19m were overexpressed in *E. coli* BL21(DE3) as follows. An overnight LB 50 µg/mL kanamycin preculture was used to inoculate 10 mL LB cultures to an OD_{600nm} of 0.1 for DfrB-H, DfrB-H-Seg and their controls. Five mL cultures of DfrB-H and their controls were incubated at 37°C for 3 h, followed by overnight 1 mM IPTG induction at 30 °C, 230 rpm. Cultures of DfrB-H-Seg and their controls were incubated at 37 °C for 3 h, followed by 3 h 1mM IPTG induction at 37 °C, 230 rpm. As for the other DfrB proteins, 10 mL culture were prepared in autoinduction media ZYP-5052 (928 mL ZY (1% tryptone, 0.5% yeast extract), 50 mL 20 x P (50 mM Na₂HPO₄, 50 mM KH₂PO₄, 25 mM (NH₄)₂SO₄), 20 mL 50 x 5052 (0.5% glycerol, 0.05% glucose, 0.2% α-lactose), 2 mL MgSO₄ (2 mM) and 0.2 mL 1000 x trace elements (0.2x)) and incubated at 37°C up to an OD_{600nm} of 0.1, followed by an overnight induction at 22°C, 230 rpm. The cultures were centrifuged at 12,800g for 30 min at 21 °C and the pellets were stored at -72°C. Cell pellets were thawed at room temperature (RT) for 30 min and resuspended in 600 µL of lysis buffer (0.1 M potassium phosphate buffer pH 8, 10 mM MgSO₄ (Anachemia), 1 mM dithiothreitol (Fisher), 0.5 mg/mL lysozyme (MP Biomedicals), 0.4 U DNase (Thermo), 1.5 mM benzamidine (Fisher), 0.25 mM phenylmethylsulfonyl fluoride (Bioshop). Cells were incubated at RT for 2h with vigorous shaking. Following centrifugation at 20,800g for 30 min at 21°C, 100 µL of the respective supernatants were transferred to 0.2 mL flat cap PCR (Fisher) tubes and incubated at different temperatures for 10 min, followed by 10 min on ice. The heated lysates were centrifuged at 12,800g for 15 min at 21°C. The supernatants were resolved in 10% tricine-SDS-PAGE gels to visualise the protein content of each variant incubated at various temperatures.

DHF (synthesised as previously reported⁷²) and NADPH (Chem Impex) were quantified spectrophotometrically in 50 mM pH 7 potassium phosphate buffer ($\epsilon^{\text{DHF}}_{282\text{nm}}$ 28,400 M⁻¹cm⁻¹ and

$\epsilon^{\text{NADPH}}_{340\text{nm}}$ 6,200 M⁻¹cm⁻¹). In a 96-well UV transparent plate (Corning), 10 μL of lysate was added to 100 μM NADPH and 100 μM DHF in 50 mM potassium phosphate buffer pH 7 for a final volume of 100 μL . Enzyme activity was determined by monitoring the depletion of DHF and NADPH at 340 nm with a plate reader (Beckman Coulter DTX 880). Initial rate of the reaction was determined on the first 20% of reaction (substrate conversion to product) with the depletion of NADPH and DHF at 340 nm ($\Delta\epsilon_{340\text{nm}}$ 12300 M⁻¹cm⁻¹ to determine product formation). Assays were carried out in triplicate.

4.5.5 Protein expression and purification

Expression of DfrB1, DfrB-H and DfrB-H5-Seg transformed in *E. coli* BL21(DE3) was carried out as follows. Overnight precultures of 5 mL inoculated 500 mL of Terrific Broth (TB) medium containing 50 $\mu\text{g}/\text{mL}$ kanamycin (Sigma). After initial growth at 37°C up to OD_{600nm} of 0.6, cells were induced by 1 mM IPTG (Thermo) and expression was carried out either at 30°C or 37°C overnight, with the exception of DfrB-H5-Seg (expression for 3 h only). The cells were harvested and resuspended in IMAC A buffer (600 mM NaCl, 50 mM Tris, 1 mM CaCl₂, 20 mM Imidazole, pH 8), lysed with a cell disrupter (Constant Systems) and centrifuged at 16,000g (Sorvall SLA-3000) at 4°C for 30 min. The supernatant was then either filtered with a 0.2 μm filter, injected onto a HisTrap FF column (Cytiva) and eluted with IMAC B buffer (600 mM NaCl, 50 mM Tris, 1 mM CaCl₂, 500 mM Imidazole, pH 8), or directly incubated with Ni-Profinity IMAC Resin (BioRad), eluted using IMAC B buffer and further purified using a Superdex 75 column (1.6 \times 55 cm) equilibrated with 50 mM pH 8 potassium phosphate buffer. Buffer exchange and concentration of protein fractions were carried out with Amicon Ultra Centrifugal Filter Units of 3K or 30K molecular weight cut-offs (Fisher).

Expression and purification of DfrB1 without a His₆-tag tag in *E. coli* BL21(DE3) was performed as follows. An overnight preculture was used to inoculate 600 mL (3 \times 200 mL) of auto-induction ZYP-5052 medium containing 50 $\mu\text{g}/\text{mL}$ kanamycin. The culture was incubated at 37°C for 3 h, followed by an overnight induction at 22°C. The cells were harvested and resuspended in 30 mL of pH 8 potassium phosphate buffer, lysed with a cell disruptor and centrifuged at 16,000g at 4°C for 25 min. The supernatant was heated at 75°C for 10 min, then cooled on ice for 10 min. The heated lysate was centrifuged at 12,800g at 4°C for 20 min. The supernatant was concentrated with an Amicon Filter, filtered and injected into the Superdex 75 column. Pure fractions were pooled together and concentrated.

The mass of each purified protein was confirmed by the Regional Mass Spectrometry Centre at Université de Montréal.

4.5.6 Kinetic parameters K_M and k_{cat}

DHF and NADPH were quantified as described in the section Dihydrofolate reductase activity in lysate. Kinetic assays were performed in a 1-cm pathlength quartz cuvette at 27°C in a Cary 100 Bio UV-Visible (Agilent) spectrophotometer by monitoring the initial rate of linear depletion of NADPH and DHF at 340 nm ($\Delta\epsilon_{340nm}$ 12,300 M⁻¹cm⁻¹ ⁷³) in 50 mM pH 7 potassium phosphate buffer. For the determination of K_M^{DHF} and K_M^{NADPH} , the concentration range of the variable substrate spanned from 3 to 145 μ M. The second substrate was kept at a saturating concentration of 50 μ M, except for DfrB1 W38F and DfrB-H5 W236F for which it was kept at 100 μ M. Data were fit to the Michaelis-Menten equation using non-linear regression analysis, with the exception of DfrB1 W38F which was fitted to the Lineweaver-Burk representation, using GraphPad Prism version 7 for Mac (GraphPad Software, San Diego, CA). Standard deviation is shown.

4.5.7 Inhibition assays

Inhibitor **1** (2,2'-[1,5-pentanediy]bis(4-oxyphenylene)]-bis-1H-benzimidazole-5-carboxylic acid) ⁴⁶ and **2** (2-Amino-N-(2-(4-((((2S,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-3,4-dihydroxytetrahydrofuran-2-yl)methyl)thio)piperidin-1-yl)ethyl)-7,7-dimethyl-4-oxo-3,4,7,8-tetrahydropteridine-6-carboxamide) ⁷⁴ were dissolved in DMSO and 50 mM pH 7 potassium phosphate buffer respectively, to prepare stocks of 10 mM. Both inhibitors were subsequently diluted to the appropriate concentrations for the inhibition assay (0 – 400 μ M). The inhibition assay of DfrB-H5 consisted of 50 μ M DHF and 50 μ M NADPH in a final volume of 100 μ L of 50 mM pH 7 potassium phosphate buffer, in addition of the diluted inhibitor. The inhibition assay of **1** was performed in 10% DMSO. The reaction was initiated by adding ~0.006 mg of purified DfrB-H5 to the reaction mix. The detection of enzyme activity was described above. The IC50 values were determined with GraphPad Prism using the log[inhibitor] versus response (four parameters) equation, using a 95% confidence interval to establish error. The K_i constant for the respective substrates were calculated using the Cheng-Prusoff equation ⁷⁵:

$$K_i = \frac{IC_{50}}{1 + \frac{[NADPH]}{K_M^{NADPH}}}$$

4.5.8 Negative-stain electron microscopy sample preparation

Purified protein from a single and homogenous gel filtration peak was diluted to a concentration of 40 ng/ μ L. Sample (3 μ L) was applied on a glow-discharged 300 mesh copper grid for 1 min and negatively stained with three consecutive droplets of 0.75% (w/v) uranyl formate solution (Electron Microscopy Sciences). The grid was blotted with Whatman filter paper to remove staining excess and air-dried at room temperature.

4.5.9 Electron microscopy data collection and analysis

Data was acquired using FEI Tecnai T12 120 kV TEM equipped with a FEI Eagle 4k x 4k CCD camera at a magnification of 67,000 × with a pixel size of 1.65 Å for DfrB-H5 and a magnification of 110,000 × with a pixel size of 0.98 Å for DfrB1. Each image was acquired using a one second exposure time with a total dose of 50 e⁻/Å² and a defocus of -1.3 μm. 2D classification was performed using cryoSPARC⁷⁶. CTFFIND4 (Wrapper) was used for the contrast transfer function estimation⁷⁷.

4.5.10 Size exclusion chromatography-multi-angle laser light scattering

Absolute molar mass was calculated using the ÄKTAmicro system (GE Healthcare) coupled with a Dawn HELEOS II MALLS detector and an OptiLab T-rEX online refractive index detector (Wyatt Technology). 500 μL of protein sample (3.27 mg/mL for DfrB1 and 0.57 mg/mL for DfrB-H5) was injected onto the Superdex 200 10/300 GL HPLC size-exclusion column (Cytiva) for DfrB-H5 and Superdex 75 10/300 GL (Cytiva) for DfrB1 at a flow rate of 0.4 mL min⁻¹. BSA was used for calibration.

4.5.11 Native mass spectrometry

DfrB1 WT and its W38F mutant were buffer exchanged into 200 mM ammonium acetate (MS grade) pH 7.5 using Micro Bio-Spin 6 columns (Bio-Rad). Platinum coated borosilicate nanospray emitters were prepared in-house as described previously⁷⁸, and were used to electrospray 10 μM of buffer-exchanged proteins on a Synapt G2-Si ion mobility mass spectrometer (Waters) equipped with a nanospray ESI source. MS data for both proteins were acquired in positive ion and sensitivity modes using the following instrumental settings: capillary voltage = 1.5 kV, cone voltage = 50 V, source offset = 50 V, source temperature = 65°C, trap gas (argon) flow = 2 mL/min and trap DC bias = 35 V. For on mobility measurements the following settings were used: IMS gas (nitrogen) flow = 90 mL/min, IMS bias = 3 V, IMS wave velocity = 550 m/s and wave height = 40 V. Data were collected in triplicate for both proteins.

4.5.12 Native gel

The quaternary structure of purified proteins was analysed with clear native polyacrylamide gel electrophoresis (CN-PAGE) using pH 7, 4% to 16% Bis-Tris NativePage gels (Thermo) as previously described⁷⁹. Briefly, protein samples were prepared in a loading dye composed of 50 mM Bis-Tris, 5% bromophenol blue (Fisher) 500 mM 6-aminocaproic acid and 10% glycerol. Following loading, native electrophoresis was performed at 4°C using a cathode buffer composed of 50 mM tricine and 15 mM Bis-Tris, pH 7, and an anode buffer of 50 mM Bis-Tris, pH 7. Following migration, gels were stained with Coomassie brilliant blue R-250 and destained with 10% acetic acid and 45% methanol.

4.5.13 Size exclusion chromatography

The oligomerisation states of DfrB-H5 WT and W236F were analysed using analytical SEC with an ÄKTA FPLC system. The 2.4 mL size exclusion column (Superdex 200 Increase 3.2/300, Cytiva) was calibrated with the Cytiva Gel Filtration Calibration Kit. 2.5 mg/mL protein solutions (10 µL injections) were applied onto the column equilibrated with 50 mM potassium phosphate, pH 8, at a flow rate of 0.075 mL/min.

4.5.14 Circular dichroism

Far-UV (250–200 nm) CD protein spectra were recorded at 20°C and 95°C using a JASCO J-815 spectropolarimeter, in a 1 mm optical path length cuvette. Denaturation spectra were performed at 220 nm, heating from 20 to 95°C with an increase of 1°C/min. Protein samples (40 µM unless otherwise indicated) were prepared in 200 µL of 50 mM potassium phosphate buffer (pH 8, unless otherwise indicated). The temperature was regulated using a Peltier-type JASCO CDF-426S/15 thermostatic controller. All experiments were performed in triplicate. The data were corrected from the background and extracted using the Spectra Manager Suite (JASCO). Mean values and standard error of the CD spectra were analysed and plotted using GraphPad Prism 7.0. The denaturation graphs present all data points from the triplicate measurements.

4.5.15 Thermotolerance assay

The initial rate of the reaction for each protein was independently determined for the various incubation temperatures. The relative activity (RA) for each enzyme was determined by comparing the initial rate of each temperature to the initial rate of the reaction at RT ($RA (\%) = \text{initial rate incubated } T^{\circ}\text{C} / \text{initial rate at } 22^{\circ}\text{C} * 100$). Data were analysed with GraphPad Prism 9.

4.6 Authors' contributions

C.L.-S.-D.: conceptualization, investigation, methodology, visualization, writing—original draft, writing—review and editing; L.A.: conceptualization, investigation, writing—review and editing; Z.J.: investigation, methodology, visualization, writing—review and editing; K.L.: investigation, methodology, visualization, writing—review and editing; M.S.-A.: investigation, writing—review and editing; K.H.: investigation; D.V.: investigation, writing—review and editing; N.W.W.: investigation, visualization, writing—review and editing; M.L.: methodology, writing—review and editing; C.J.T.: funding acquisition, writing—review and editing; N.D.: funding acquisition, writing—review and editing; C.B.: funding acquisition, writing—review and editing; J.N.C.: conceptualization, investigation, writing—review and editing; J.N.P.: conceptualization, funding acquisition, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

4.7 Funding

This work was supported by NSERC discovery grants RGPIN-N-2018-04686 (J.N.P.) and RGPIN-2022-04368 (N.D.), the Canada Research Chair in Engineering of Applied Proteins (J.N.P.) and FRQ-S Research Scholar Senior Career Award 281993 (N.D.). C.L.-S.-D. and L.A. are grateful to NSERC, FQRNT and Université de Montréal for scholarships. K.L. is supported by a CREATE- APRENTICE scholarship.

4.8 Acknowledgement

This research is dedicated to the memory of Elizabeth E. Howell, foremost expert in DfrB structure and function. We thank Normand Cyr for his help with SEC-MALLS data analysis and Samy Cecioni for providing access to instruments. We thank Stella Cellier-Goetghebeur and Nobuhiko Tokuriki for proofreading.

4.9 Conflict of interest declaration

We declare we have no competing interests.

4.10 References

- (1) Knox, J. R.; Moews, P. C.; Frere, J.-M. Molecular Evolution of Bacterial β -Lactam Resistance. *Chem. Biol.* **1996**, *3* (11), 937–947. [https://doi.org/10.1016/S1074-5521\(96\)90182-9](https://doi.org/10.1016/S1074-5521(96)90182-9).
- (2) Palzkill, T. Structural and Mechanistic Basis for Extended-Spectrum Drug-Resistance Mutations in Altering the Specificity of TEM, CTX-M, and KPC β -Lactamases. *Front. Mol. Biosci.* **2018**, *5*, 16. <https://doi.org/10.3389/fmolb.2018.00016>.
- (3) Noor, S.; Taylor, M. C.; Russell, R. J.; Jermin, L. S.; Jackson, C. J.; Oakeshott, J. G.; Scott, C. Intramolecular Epistasis and the Evolution of a New Enzymatic Function. *PLoS ONE* **2012**, *7* (6), e39822. <https://doi.org/10.1371/journal.pone.0039822>.
- (4) Yang, G.; Anderson, D. W.; Baier, F.; Dohmen, E.; Hong, N.; Carr, P. D.; Kamerlin, S. C. L.; Jackson, C. J.; Bornberg-Bauer, E.; Tokuriki, N. Higher-Order Epistasis Shapes the Fitness Landscape of a Xenobiotic-Degrading Enzyme. *Nat. Chem. Biol.* **2019**, *15* (11), 1120–1128. <https://doi.org/10.1038/s41589-019-0386-3>.
- (5) Copley, S. D. Evolution of Efficient Pathways for Degradation of Anthropogenic Chemicals. *Nat. Chem. Biol.* **2009**, *5* (8), 559–566. <https://doi.org/10.1038/nchembio.197>.
- (6) Jensen, R. A. Enzyme Recruitment in Evolution of New Function. *Annu. Rev. Microbiol.* **1976**, *30* (1), 409–425. <https://doi.org/10.1146/annurev.mi.30.100176.002205>.
- (7) Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Gould, S. M.; Roodveldt, C.; Tawfik, D. S. The “evolvability” of Promiscuous Protein Functions. *Nat. Genet.* **2005**, *37* (1), 73–76. <https://doi.org/10.1038/ng1482>.
- (8) Tawfik, O. K. and D. S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annu. Rev. Biochem.* **2010**, *79* (1), 471–505. <https://doi.org/10.1146/annurev-biochem-030409-143718>.

- (9) Copley, S. D. An Evolutionary Biochemist's Perspective on Promiscuity. *Trends Biochem. Sci.* **2015**, *40* (2), 72–78. <https://doi.org/10.1016/j.tibs.2014.12.004>.
- (10) Kaltenbach, M.; Tokuriki, N. Dynamics and Constraints of Enzyme Evolution. *J. Exp. Zoolog. B Mol. Dev. Evol.* **2014**, *322* (7), 468–487. <https://doi.org/10.1002/jez.b.22562>.
- (11) Crean, R. M.; Gardner, J. M.; Kamerlin, S. C. L. Harnessing Conformational Plasticity to Generate Designer Enzymes. *J. Am. Chem. Soc.* **2020**, *142* (26), 11324–11342. <https://doi.org/10.1021/jacs.0c04924>.
- (12) Campbell, E.; Kaltenbach, M.; Correy, G. J.; Carr, P. D.; Porebski, B. T.; Livingstone, E. K.; Afriat-Jurnou, L.; Buckle, A. M.; Weik, M.; Hollfelder, F.; Tokuriki, N.; Jackson, C. J. The Role of Protein Dynamics in the Evolution of New Enzyme Function. *Nat. Chem. Biol.* **2016**, *12* (11), 944–950. <https://doi.org/10.1038/nchembio.2175>.
- (13) Zou, T.; Risso, V. A.; Gavira, J. A.; Sanchez-Ruiz, J. M.; Ozkan, S. B. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Mol. Biol. Evol.* **2015**, *32* (1), 132–143. <https://doi.org/10.1093/molbev/msu281>.
- (14) James, L. C.; Tawfik, D. S. Conformational Diversity and Protein Evolution – a 60-Year-Old Hypothesis Revisited. *Trends Biochem. Sci.* **2003**, *28* (7), 361–368. [https://doi.org/10.1016/S0968-0004\(03\)00135-X](https://doi.org/10.1016/S0968-0004(03)00135-X).
- (15) Khersonsky, O.; Roodveldt, C.; Tawfik, D. Enzyme Promiscuity: Evolutionary and Mechanistic Aspects. *Curr. Opin. Chem. Biol.* **2006**, *10* (5), 498–508. <https://doi.org/10.1016/j.cbpa.2006.08.011>.
- (16) Kirby, R. Evolutionary Origin of the Class A and Class C β -Lactamases. *J. Mol. Evol.* **1992**, *34* (4), 345–350. <https://doi.org/10.1007/BF00160242>.
- (17) Pratt, R. F. β -Lactamases: Why and How: Miniperspective. *J. Med. Chem.* **2016**, *59* (18), 8207–8220. <https://doi.org/10.1021/acs.jmedchem.6b00448>.
- (18) Noda-Garcia, L.; Tawfik, D. S. Enzyme Evolution in Natural Products Biosynthesis: Target- or Diversity-Oriented? *Curr. Opin. Chem. Biol.* **2020**, *59*, 147–154. <https://doi.org/10.1016/j.cbpa.2020.05.011>.
- (19) Todd, A. E.; Orengo, C. A.; Thornton, J. M. Sequence and Structural Differences between Enzyme and Nonenzyme Homologs. *Structure* **2002**, *10* (10), 1435–1451. [https://doi.org/10.1016/S0969-2126\(02\)00861-4](https://doi.org/10.1016/S0969-2126(02)00861-4).
- (20) Tam, R.; Saier, M. H. A Bacterial Periplasmic Receptor Homologue with Catalytic Activity: Cyclohexadienyl Dehydratase of *Pseudomonas Aeruginosa* Is Homologous to Receptors Specific for Polar Amino Acids. *Res. Microbiol.* **1993**, *144* (3), 165–169. [https://doi.org/10.1016/0923-2508\(93\)90041-Y](https://doi.org/10.1016/0923-2508(93)90041-Y).
- (21) Ngaki, M. N.; Louie, G. V.; Philippe, R. N.; Manning, G.; Pojer, F.; Bowman, M. E.; Li, L.; Larsen, E.; Wurtele, E. S.; Noel, J. P. Evolution of the Chalcone-Isomerase Fold from Fatty-Acid Binding to Stereospecific Catalysis. *Nature* **2012**, *485* (7399), 530–533. <https://doi.org/10.1038/nature11009>.

- (22) Ortmayer, M.; Lafite, P.; Menon, B. R. K.; Tralau, T.; Fisher, K.; Denkhaus, L.; Scrutton, N. S.; Rigby, S. E. J.; Munro, A. W.; Hay, S.; Leys, D. An Oxidative N-Demethylase Reveals PAS Transition from Ubiquitous Sensor to Enzyme. *Nature* **2016**, *539* (7630), 593–597. <https://doi.org/10.1038/nature20159>.
- (23) Kaltenbach, M.; Burke, J. R.; Dindo, M.; Pabis, A.; Munsberg, F. S.; Rabin, A.; Kamerlin, S. C. L.; Noel, J. P.; Tawfik, D. S. Evolution of Chalcone Isomerase from a Noncatalytic Ancestor. *Nat. Chem. Biol.* **2018**, *14* (6), 548–555. <https://doi.org/10.1038/s41589-018-0042-3>.
- (24) Clifton, B. E.; Kaczmarek, J. A.; Carr, P. D.; Gerth, M. L.; Tokuriki, N.; Jackson, C. J. Evolution of Cyclohexadienyl Dehydratase from an Ancestral Solute-Binding Protein. *Nat. Chem. Biol.* **2018**, *14* (6), 542–547. <https://doi.org/10.1038/s41589-018-0043-2>.
- (25) Wahba, H. M.; Lecoq, L.; Stevenson, M.; Mansour, A.; Cappadocia, L.; Lafrance-Vanasse, J.; Wilkinson, K. J.; Sygusch, J.; Wilcox, D. E.; Omichinski, J. G. Structural and Biochemical Characterization of a Copper-Binding Mutant of the Organomercurial Lyase MerB: Insight into the Key Role of the Active Site Aspartic Acid in Hg–Carbon Bond Cleavage and Metal Binding Specificity. *Biochemistry* **2016**, *55* (7), 1070–1081. <https://doi.org/10.1021/acs.biochem.5b01298>.
- (26) Kaur, G.; Subramanian, S. Repurposing TRASH: Emergence of the Enzyme Organomercurial Lyase from a Non-Catalytic Zinc Finger Scaffold. *J. Struct. Biol.* **2014**, *188* (1), 16–21. <https://doi.org/10.1016/j.jsb.2014.09.001>.
- (27) Lemay-St-Denis, C.; Diwan, S.-S.; Pelletier, J. N. The Bacterial Genomic Context of Highly Trimethoprim-Resistant DfrB Dihydrofolate Reductases Highlights an Emerging Threat to Public Health. *Antibiotics* **2021**, *10* (4), 433. <https://doi.org/10.3390/antibiotics10040433>.
- (28) Eliopoulos, G. M.; Huovinen, P. Resistance to Trimethoprim-Sulfamethoxazole. *Clin. Infect. Dis.* **2001**, *32* (11), 1608–1614. <https://doi.org/10.1086/320532>.
- (29) World Health Organization. WHO Report on Surveillance of Antibiotic Consumption: 2016–2018 Early Implementation, 2018.
- (30) Cuong, N. V.; Padungtod, P.; Thwaites, G.; Carrique-Mas, J. J. Antimicrobial Usage in Animal Production: A Review of the Literature with a Focus on Low- and Middle-Income Countries. *Antibiot. Basel Switz.* **2018**, *7* (3). <https://doi.org/10.3390/antibiotics7030075>.
- (31) Pohl, E.; Holmes, R. K.; Hol, W. G. J. Crystal Structure of a Cobalt-Activated Diphtheria Toxin Repressor-DNA Complex Reveals a Metal-Binding SH3-like Domain. *J. Mol. Biol.* **1999**, *292* (3), 653–667. <https://doi.org/10.1006/jmbi.1999.3073>.
- (32) Kishan, K.; Agrawal, V. SH3-like Fold Proteins Are Structurally Conserved and Functionally Divergent. *Curr. Protein Pept. Sci.* **2005**, *6* (2), 143–150. <https://doi.org/10.2174/1389203053545444>.
- (33) León, E.; Navarro-Avilés, G.; Santiveri, C. M.; Flores-Flores, C.; Rico, M.; González, C.; Murillo, F. J.; Elías-Arnanz, M.; Jiménez, M. A.; Padmanabhan, S. A Bacterial Antirepressor with SH3 Domain Topology Mimics Operator DNA in Sequestering the Repressor DNA Recognition Helix. *Nucleic Acids Res.* **2010**, *38* (15), 5226–5241. <https://doi.org/10.1093/nar/gkq277>.

- (34) Eijkelenboom, A. P. A. M.; Puras Lutzke, R. A.; Boelens, R.; Plasterk, R. H. A.; Kaptein, R.; Hård, K. The DNA-Binding Domain of HIV-1 Integrase Has an SH3-like Fold. *Nat. Struct. Mol. Biol.* **1995**, *2* (9), 807–810. <https://doi.org/10.1038/nsb0995-807>.
- (35) Howell, E. E. Searching Sequence Space: Two Different Approaches to Dihydrofolate Reductase Catalysis. *ChemBioChem* **2005**, *6* (4), 590–600. <https://doi.org/10.1002/cbic.200400237>.
- (36) Reece, L. J.; Nichols, R.; Ogden, R. C.; Howell, E. E. Construction of a Synthetic Gene for an R-Plasmid-Encoded Dihydrofolate Reductase and Studies on the Role of the N-Terminus in the Protein. *Biochemistry* *30* (45), 10895–10904. <https://doi.org/10.1021/bi00109a013>.
- (37) Bar-Even, A.; Noor, E.; Savir, Y.; Liebermeister, W.; Davidi, D.; Tawfik, D. S.; Milo, R. The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **2011**, *50* (21), 4402–4410. <https://doi.org/10.1021/bi2002289>.
- (38) Toulouse, J. L.; Shi, G.; Lemay-St-Denis, C.; Ebert, M. C. C. J. C.; Deon, D.; Gagnon, M.; Ruediger, E.; Saint-Jacques, K.; Forge, D.; Vanden Eynde, J. J.; Marinier, A.; Ji, X.; Pelletier, J. N. Dual-Target Inhibitors of the Folate Pathway Inhibit Intrinsically Trimethoprim-Resistant DfrB Dihydrofolate Reductases. *ACS Med. Chem. Lett.* **2020**, *acsmedchemlett.0c00393*. <https://doi.org/10.1021/acsmedchemlett.0c00393>.
- (39) Bolin, J. T.; Filman, D. J.; Matthews, D. A.; Hamlin, R. C.; Kraut, J. Crystal Structures of Escherichia Coli and Lactobacillus Casei Dihydrofolate Reductase Refined at 1.7 Å Resolution. I. General Features and Binding of Methotrexate. *J. Biol. Chem.* **1982**, *257* (22), 13650–13662. [https://doi.org/10.1016/S0021-9258\(18\)33497-5](https://doi.org/10.1016/S0021-9258(18)33497-5).
- (40) Faltyn, M.; Alcock, B.; McArthur, A. Evolution and Nomenclature of the Trimethoprim Resistant Dihydrofolate (Dfr) Reductases. **2019**. <https://doi.org/10.20944/preprints201905.0137.v1>.
- (41) Krahn, J. M.; Jackson, M. R.; DeRose, E. F.; Howell, E. E.; London, R. E. Crystal Structure of a Type II Dihydrofolate Reductase Catalytic Ternary Complex †. *Biochemistry* **2007**, *46* (51), 14878–14888. <https://doi.org/10.1021/bi701532r>.
- (42) Strader, M. B.; Smiley, R. D.; Stinnett, L. G.; VerBerkmoes, N. C.; Howell, E. E. Role of S65, Q67, I68, and Y69 Residues in Homotetrameric R67 Dihydrofolate Reductase †. *Biochemistry* **2001**, *40* (38), 11344–11352. <https://doi.org/10.1021/bi0110544>.
- (43) Stinnett, L. G.; Smiley, R. D.; Hicks, S. N.; Howell, E. E. “Catch 222,” the Effects of Symmetry on Ligand Binding and Catalysis in R67 Dihydrofolate Reductase as Determined by Mutations at Tyr-69. *J. Biol. Chem.* **2004**, *279* (45), 47003–47009. <https://doi.org/10.1074/jbc.M404485200>.
- (44) Schmitzer, A. R.; Lépine, F.; Pelletier, J. N. Combinatorial Exploration of the Catalytic Site of a Drug-Resistant Dihydrofolate Reductase: Creating Alternative Functional Configurations. *Protein Eng. Des. Sel.* **2004**, *17* (11), 809–819. <https://doi.org/10.1093/protein/gzh090>.
- (45) West, F. W.; Seo, H.-S.; Bradrick, T. D.; Howell, E. E. Effects of Single-Tryptophan Mutations on R67 Dihydrofolate Reductase †. *Biochemistry* **2000**, *39* (13), 3678–3689. <https://doi.org/10.1021/bi992195x>.

- (46) Bastien, D.; Ebert, M. C. C. J. C.; Forge, D.; Toulouse, J.; Kadnikova, N.; Perron, F.; Mayence, A.; Huang, T. L.; Vanden Eynde, J. J.; Pelletier, J. N. Fragment-Based Design of Symmetrical Bis-Benzimidazoles as Selective Inhibitors of the Trimethoprim-Resistant, Type II R67 Dihydrofolate Reductase. *J. Med. Chem.* **2012**, *55* (7), 3182–3192. <https://doi.org/10.1021/jm201645r>.
- (47) Toulouse, J. L.; Yachnin, B. J.; Ruediger, E. H.; Deon, D.; Gagnon, M.; Saint-Jacques, K.; Ebert, M. C. C. J. C.; Forge, D.; Bastien, D.; Colin, D. Y.; Vanden Eynde, J. J.; Marinier, A.; Berghuis, A. M.; Pelletier, J. N. Structure-Based Design of Dimeric Bisbenzimidazole Inhibitors to an Emergent Trimethoprim-Resistant Type II Dihydrofolate Reductase Guides the Design of Monomeric Analogues. *ACS Omega* **2019**, *4* (6), 10056–10069. <https://doi.org/10.1021/acsomega.9b00640>.
- (48) Some, D.; Amartely, H.; Tsadok, A.; Lebendiker, M. Characterization of Proteins by Size-Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS). *J. Vis. Exp.* **2019**, No. 148, 59615. <https://doi.org/10.3791/59615>.
- (49) Nichols, R.; Weaver, C. D.; Eisenstein, E.; Blakley, R. L.; Appleman, J.; Huang, T. H.; Huang, F. Y.; Howell, E. E. Titration of Histidine 62 in R67 Dihydrofolate Reductase Is Linked to a Tetramer .Tautm. Two-Dimer Equilibrium. *Biochemistry* **1993**, *32* (7), 1695–1706. <https://doi.org/10.1021/bi00058a002>.
- (50) Hicks, S. N.; Smiley, R. D.; Hamilton, J. B.; Howell, E. E. Role of Ionic Interactions in Ligand Binding and Catalysis of R67 Dihydrofolate Reductase †. *Biochemistry* **2003**, *42* (36), 10569–10578. <https://doi.org/10.1021/bi034643d>.
- (51) Waldron, T. T.; Murphy, K. P. Stabilization of Proteins by Ligand Binding: Application to Drug Screening and Determination of Unfolding Energetics. *Biochemistry* **2003**, *42* (17), 5058–5064. <https://doi.org/10.1021/bi034212v>.
- (52) Huynh, K.; Partch, C. L. Analysis of Protein Stability and Ligand Interactions by Thermal Shift Assay. *Curr. Protoc. Protein Sci.* **2015**, *79* (1). <https://doi.org/10.1002/0471140864.ps2809s79>.
- (53) Jackson, M.; Chopra, S.; Smiley, R. D.; Maynord, P. O.; Rosowsky, A.; London, R. E.; Levy, L.; Kalman, T. I.; Howell, E. E. Calorimetric Studies of Ligand Binding in R67 Dihydrofolate Reductase. *Biochemistry* **2005**, *44* (37), 12420–12433. <https://doi.org/10.1021/bi050881s>.
- (54) Ebert, M. C. C. J. C.; Morley, K. L.; Volpato, J. P.; Schmitzer, A. R.; Pelletier, J. N. Asymmetric Mutations in the Tetrameric R67 Dihydrofolate Reductase Reveal High Tolerance to Active-Site Substitutions: Asymmetric Mutations in R67 Dihydrofolate Reductase. *Protein Sci.* **2015**, *24* (4), 495–507. <https://doi.org/10.1002/pro.2602>.
- (55) Pattishall, K. H.; Acar, J.; Burchall, J. J.; Goldstein, F. W.; Harvey, R. J. Two Distinct Types of Trimethoprim-Resistant Dihydrofolate Reductase Specified by R-Plasmids of Different Compatibility Groups. *J. Biol. Chem.* **1977**, *252* (7), 2319–2323.
- (56) Fling, M. E.; Richards, C. The Nucleotide Sequence of the Trimethoprim-Resistant Dihydrofolate Reductase Gene Harbored by Tn7. *Nucleic Acids Res.* **1983**, *11* (15), 5147–5158. <https://doi.org/10.1093/nar/11.15.5147>.

- (57) Amyes, S. G. B.; Smith, J. T. R-Factor Trimethoprim Resistance Mechanism: An Insusceptible Target Site. *Biochem. Biophys. Res. Commun.* **1974**, *58* (2), 412–418. [https://doi.org/10.1016/0006-291X\(74\)90380-5](https://doi.org/10.1016/0006-291X(74)90380-5).
- (58) Deng, H.; Callender, R.; Howell, E. Vibrational Structure of Dihydrofolate Bound to R67 Dihydrofolate Reductase. *J. Biol. Chem.* **2001**, *276* (52), 48956–48960. <https://doi.org/10.1074/jbc.M105107200>.
- (59) Wan, Q.; Bennett, B. C.; Wilson, M. A.; Kovalevsky, A.; Langan, P.; Howell, E. E.; Dealwis, C. Toward Resolving the Catalytic Mechanism of Dihydrofolate Reductase Using Neutron and Ultrahigh-Resolution X-Ray Crystallography. *Proc. Natl. Acad. Sci.* **2014**, *111* (51), 18225–18230. <https://doi.org/10.1073/pnas.1415856111>.
- (60) The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuhe, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L.-S.; Zhang, J.; Ruch, P.; Teodoro, D. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>.
- (61) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D412–D419. <https://doi.org/10.1093/nar/gkaa913>.
- (62) Madeira, F.; Pearce, M.; Tivey, A. R. N.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R. Search and Sequence Analysis Tools Services from EMBL-EBI in 2022. *Nucleic Acids Res* **2022**, gkac240. <https://doi.org/10.1093/nar/gkac240>.
- (63) Trifinopoulos, J.; Nguyen, L.-T.; von Haeseler, A.; Minh, B. Q. W-IQ-TREE: A Fast Online Phylogenetic Tool for Maximum Likelihood Analysis. *Nucleic Acids Res.* **2016**, *44* (W1), W232–W235. <https://doi.org/10.1093/nar/gkw256>.

- (64) Letunic, I.; Bork, P. Interactive Tree of Life (ITOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees. *Nucleic Acids Res.* **2016**, *44* (W1), W242-245. <https://doi.org/10.1093/nar/gkw290>.
- (65) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. *ColabFold - Making Protein Folding Accessible to All*; preprint; Bioinformatics, 2021. <https://doi.org/10.1101/2021.08.15.456425>.
- (66) Notredame, C.; Higgins, D. G.; Heringa, J. T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment 1 Edited by J. Thornton. *J. Mol. Biol.* **2000**, *302* (1), 205–217. <https://doi.org/10.1006/jmbi.2000.4042>.
- (67) Nguyen, L.-T.; Schmidt, H. A.; von Haeseler, A.; Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32* (1), 268–274. <https://doi.org/10.1093/molbev/msu300>.
- (68) Clouthier, C. M.; Morin, S.; Gobeil, S. M. C.; Doucet, N.; Blanchet, J.; Nguyen, E.; Gagné, S. M.; Pelletier, J. N. Chimeric β -Lactamases: Global Conservation of Parental Function and Fast Time-Scale Dynamics with Increased Slow Motions. *PLoS ONE* **2012**, *7* (12), e52283. <https://doi.org/10.1371/journal.pone.0052283>.
- (69) van den Ent, F.; Löwe, J. RF Cloning: A Restriction-Free Method for Inserting Target Genes into Plasmids. *J. Biochem. Biophys. Methods* **2006**, *67* (1), 67–74. <https://doi.org/10.1016/j.jbbm.2005.12.008>.
- (70) Bond, S. R.; Naus, C. C. RF-Cloning.Org: An Online Tool for the Design of Restriction-Free Cloning Projects. *Nucleic Acids Res.* **2012**, *40* (W1), W209–W213. <https://doi.org/10.1093/nar/gks396>.
- (71) Wiegand, I.; Hilpert, K.; Hancock, R. E. W. Agar and Broth Dilution Methods to Determine the Minimal Inhibitory Concentration (MIC) of Antimicrobial Substances. *Nat. Protoc.* **2008**, *3* (2), 163–175. <https://doi.org/10.1038/nprot.2007.521>.
- (72) Blakley, R. L. Crystalline Dihydropteroylglutamic Acid. *Nature* **1960**, *188* (4746), 231–232. <https://doi.org/10.1038/188231a0>.
- (73) Baccanari, D.; Phillips, A.; Smith, S.; Sinski, D.; Burchall, J. Purification and Properties of Escherichia Coli Dihydrofolate Reductase. *Biochemistry* **1975**, *14* (24), 5267–5273. <https://doi.org/10.1021/bi00695a006>.
- (74) Shi, G.; Shaw, G.; Liang, Y.-H.; Subburaman, P.; Li, Y.; Wu, Y.; Yan, H.; Ji, X. Bisubstrate Analogue Inhibitors of 6-Hydroxymethyl-7,8-Dihydropterin Pyrophosphokinase: New Design with Improved Properties. *Bioorg. Med. Chem.* **2012**, *20* (1), 47–57. <https://doi.org/10.1016/j.bmc.2011.11.032>.
- (75) Yung-Chi, C.; Prusoff, W. H. Relationship between the Inhibition Constant (KI) and the Concentration of Inhibitor Which Causes 50 per Cent Inhibition (I50) of an Enzymatic Reaction. *Biochem. Pharmacol.* **1973**, *22* (23), 3099–3108. [https://doi.org/10.1016/0006-2952\(73\)90196-2](https://doi.org/10.1016/0006-2952(73)90196-2).

- (76) Punjani, A.; Rubinstein, J. L.; Fleet, D. J.; Brubaker, M. A. CryoSPARC: Algorithms for Rapid Unsupervised Cryo-EM Structure Determination. *Nat. Methods* **2017**, *14* (3), 290–296. <https://doi.org/10.1038/nmeth.4169>.
- (77) Rohou, A.; Grigorieff, N. CTFFIND4: Fast and Accurate Defocus Estimation from Electron Micrographs. *J. Struct. Biol.* **2015**, *192* (2), 216–221. <https://doi.org/10.1016/j.jsb.2015.08.008>.
- (78) Weerasinghe, N. W.; Habibi, Y.; Uggowitz, K. A.; Thibodeaux, C. J. Exploring the Conformational Landscape of a Lanthipeptide Synthetase Using Native Mass Spectrometry. *Biochemistry* **2021**, *60* (19), 1506–1519. <https://doi.org/10.1021/acs.biochem.1c00085>.
- (79) Charbonneau, D. M.; Aubé, A.; Rachel, N. M.; Guerrero, V.; Delorme, K.; Breault-Turcot, J.; Masson, J.-F.; Pelletier, J. N. Development of *Escherichia Coli* Asparaginase II for Immunosensing: A Trade-Off between Receptor Density and Sensing Efficiency. *ACS Omega* **2017**, *2* (5), 2114–2125. <https://doi.org/10.1021/acsomega.7b00110>.

4.11 Supplementary material

Table S4.1. Taxonomic information on the DfrB-H characterized in this study

	UniProtKB ID	Sequence length (residues)	Superkindom	Class	Order	Host organism
DfrB-H2	Q4VCS1	97	Bacteria	γ - <i>proteobacteria</i>	<i>Enterobacterales</i>	<i>Klebsiella pneumoniae</i>
DfrB-H3	A0A0H5ARK7	167	Virus			<i>Pseudomonas phage PS-1</i>
DfrB-H4	A0A2I7QNH6	218	Virus			<i>Vibrio phage</i>
DfrB-H5	A0A1I4V6W3	365	Bacteria	α - <i>proteobacteria</i>	<i>Hyphomicrobiales</i>	<i>Methylobacterium pseudosasicola</i>
DfrB-H6	A0A087N9R8	463	Bacteria	α - <i>proteobacteria</i>	<i>Sphingomonadales</i>	<i>Sphingobium sp. ba1</i>

Table S4.2. MICs performed in IPTG induction broth.

Values differ slightly from the agar method.

DfrB1	MIC [TMP] µg/mL	
	DfrB-H	DfrB-H-Seg
	600	
2	600	600
3	4.69	4.69
4	600	300
5	600	>600
6	300	600

Table S4.3. MICs for deleterious mutations in DfrB1 and DfrB-H5.

Assay performed on agar IPTG induction media.

DfrB1 W38F	DfrB-H5 W236F	DfrB1 Y69L	DfrB-H5 Y267L



Figure S4.1. Sequence alignment of representative bacterial FoIA and DfrA.

Important active-site residues are underlined and in bold. The alignment was performed with Clustal Omega.

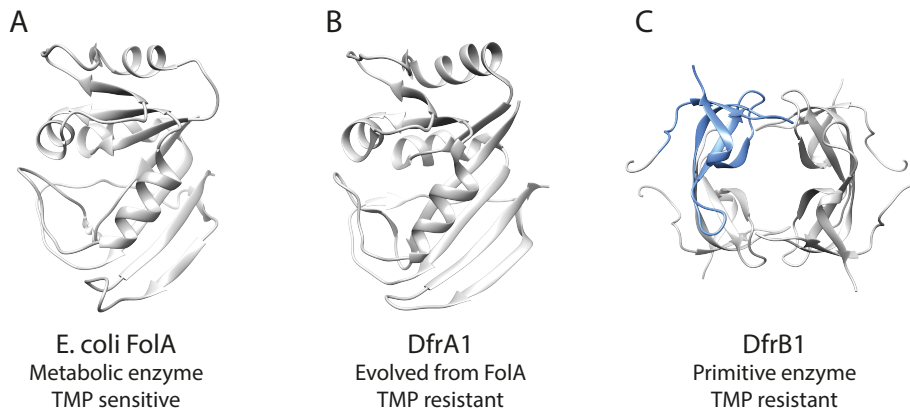


Figure S4.2. Structure of characterized dihydrofolate reductases.

A. *E. coli* FoaA (PDB 1DDR). Monomeric, 159 amino acids. **B.** DfrA1, a type A dihydrofolate reductase (PDB 5ECC). Monomeric, 157 amino acids. **C.** Type B dihydrofolate reductase DfrB1 (PDB 1VIE). Homotetrameric, 78 amino acids. SH3-like domain of one monomer is colored in blue.

	N-terminus	SH3-like domain		
dfrB1	MERSSNEVSNPVAGNFVFP	SDATFGMGDRVRKKS	GAAWQGQIVGWYCTNLTPEGYAVESE 60	
dfrB2	MGQSSDEANAPVAGQFAL	PLSATFGLGDRVRKKS	GAAWQGQVVGWYCTKLTPEGYAVESE 60	
dfrB3	MDQHNGVSTLVAGQFAL	PSHATFGLGDRVRKKS	GAAWQGQVVGWYCTKLTPEGYAVESE 60	
dfrB4	MNEGKNEVSTSAAGRF	FAPSNATFALGDRVRKKS	GAAWQGRIVGWYCTTLTPEGYAVESE 60	
dfrB5	MDQGRSEVSNPVAGQF	FAPSNAAF	FGMGDRVRKKS	GAAWQGQIVGWYCTKLTPEGYAVESE 60
dfrB6	MDQGSNEVINPVAGQF	ASPSNATF	FGMGDRVRKKS	GAAWQGQIVGWYSTKLTPEGYAVESE 60
dfrB7	MDQGSNEVGNPVAGQF	SFSPSNA	AFSMGDRVRKKS	GAAWQGQIVGWYCTKLTPEGYAVESE 60
dfrB9	MNQSSNCISTPVVGQF	ALPFQPTFGLGDRVRKKS	GAAWQGQVVGWYCTKLTPEGYAVESE 60	
dfrB10	MDQSSNEVSTPVAGQF	ALPLRATFGLGDRVRKKS	GAAWQGQVVGWYCTKLTPEGYAVESE 60	
dfrB11	MDQSSKEVGT	PVVGQFALPSHATFGLGDRVRKKS	GAAWQGQVVGWYCTKLTPEGYAVESE 60	
	** * * :* :***** :***** :*****		
dfrB1	AHPGS	<u>VQIY</u> PVAALERIN----	78	
dfrB2	SHPGS	<u>VQIY</u> PVAALERVA----	78	
dfrB3	SHPGS	<u>VQIY</u> PMTALERVA----	78	
dfrB4	SHPGS	<u>VQIY</u> PVAALERVA----	78	
dfrB5	AHPGS	<u>VQIY</u> PVAALERIN----	78	
dfrB6	AHPGS	<u>VQIY</u> PVAALERNV----	78	
dfrB7	AHPGS	<u>VQIY</u> PVAALERINGVQG	82	
dfrB9	AHPGS	<u>VQIY</u> PVAALERVA----	78	
dfrB10	SHPGS	<u>VQIY</u> PVAALERVA----	78	
dfrB11	SHPGS	<u>VQIY</u> PVNALERVA----	78	
	:*****:	****:		

Figure S4.3. Sequence alignment of all characterized DfrB.

The N-terminal domain and the five beta strands of the SH3-like fold are indicated with blue arrows. Active-site residues are underlined and in bold. The alignment was performed with Clustal Omega.

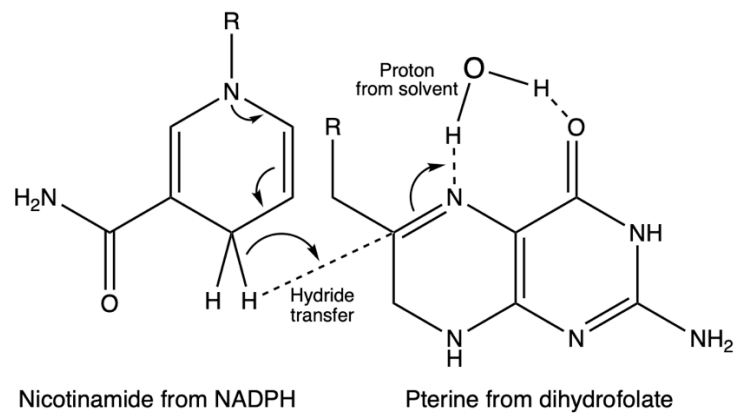


Figure S4.4. Catalytic mechanism for dihydrofolate reduction by DfrB1.
Figure adapted from ¹.

P00383_(DfrB1)	-----	70
Q4VCS1_(DfrB-H2)	-----MQRVVGPHRTPRSSQ-----	70
P00384_(DfrB2)	-----	70
Q71RY0_(DfrB3)	-----	70
B0FMU1_(DfrB4)	-----	70
A0A0U2TX79_(DfrB5)	-----	70
Q1HVJ4_(DfrB6)	-----	70
D2Y395_(DfrB7)	-----	70
R4UB13_(DfrB9)	-----	70
A0A2N2TNN4_(DfrB11)	-----	70
A0A0A0YVN9	-----	70
F4ZVS8	-----	70
A0A0A0YXA2	-----	70
A0A0A0Z1Q9	-----	70
A0A0A0YWM8	-----	70
A0A0N7AFD7	-----MQRVVGPHRTPRSSQ-----	70
A0A2I5YL54	-----	70
Q53CM2	-----	70
A0A0H5ARK7_(DfrB-H3)	M-----SEIKRYEPINIDGTCGIC-----AEDADGAYV--	70
A0A0J7Y2N3	-----	70
A0A087N9R8_(DfrB-H6)	MLVTRTRDIDCCSCGSRPRHRRIFLYRAETERIERSIRASPGGKAFHHRISIVAASEGLTMTNDKDGALLR	70
A0A2E1B2M0	M-----	70
H5WLJ1	MTR-----	70
A0A1R1QGT7	M-----	70
A0A1G7DL98	-----	70
A0A226CF51	-----	70
A0A2I7QNH6_(DfrB-H4)	MKYT-----AEQIKSILENAPSGSQFY-----TDDYDVKYR--	70
A0A0F6R7A8	MF-----ECF-----	70
A0A117UUM5	MT-----SAFY-----GPDQDAQAI--	70
A0A1I4V6W3_(DfrB-H5)	MTEH-----HTPDAASGTDR-----GEKPANIQPGELL--	70

P00383_(DfrB1)	-----	140
Q4VCS1_(DfrB-H2)	-----	140
P00384_(DfrB2)	-----	140
Q71RY0_(DfrB3)	-----	140
B0FMU1_(DfrB4)	-----	140
A0A0U2TX79_(DfrB5)	-----	140
Q1HVJ4_(DfrB6)	-----	140
D2Y395_(DfrB7)	-----	140
R4UB13_(DfrB9)	-----	140
A0A2N2TNN4_(DfrB11)	-----	140
A0A0A0YVN9	-----	140
F4ZVS8	-----	140
A0A0A0YXA2	-----	140
A0A0A0Z1Q9	-----	140
A0A0A0YWM8	-----	140
A0A0N7AFD7	-----	140
A0A2I5YL54	-----	140
Q53CM2	-----	140
A0A0H5ARK7_(DfrB-H3)	-----	140
A0A0J7Y2N3	-----	140
A0A087N9R8_(DfrB-H6)	EAIHINALSSELGDTYDCGRVWDWVHGTMGEDDFTPVNDRVGEITETVLAALQSHGQAFDAASPLGDFQC	140
A0A2E1B2M0	-----	140
H5WLJ1	-----	140
A0A1R1QGT7	-----	140
A0A1G7DL98	-----	140
A0A226CF51	-----	140
A0A2I7QNH6_(DfrB-H4)	-----	140
A0A0F6R7A8	-----	140
A0A117UUM5	-----	140
A0A1I4V6W3_(DfrB-H5)	-----W-----	140

P00383_(DfrB1)	-----	210
Q4VCS1_(DfrB-H2)	-----	210
P00384_(DfrB2)	-----	210
Q71RY0_(DfrB3)	-----	210
B0FMU1_(DfrB4)	-----	210
A0A0U2TX79_(DfrB5)	-----	210
Q1HVJ4_(DfrB6)	-----	210
D2Y395_(DfrB7)	-----	210
R4UB13_(DfrB9)	-----	210
A0A2N2TNN4_(DfrB11)	-----	210
A0A0A0YVN9	-----	210
F4ZVS8	-----	210
A0A0A0YXA2	-----	210
A0A0A0Z1Q9	-----	210
A0A0A0YWM8	-----	210

A0A0N7AFD7	-----	210
A0A2I5YL54	-----	210
Q53CM2	-----	210
A0A0H5ARK7_(DfrB-H3)	-----DYRDFAAALNDELAA--LKDGAPAQ--	210
A0A0J7Y2N3	-----MCA-----	210
A0A087N9R8_(DfrB-H6)	HNCKRIVERLTVAPEKCPDCMCCSSFSAVYGSEAKYRAAPQYAFDAAGVREAVADAIWAE--FMKMAPGYLL	210
A0A2E1B2M0	-----	210
H5WLJ1	-----	210
A0A1R1QGT7	-----	210
A0A1G7DL98	-----	210
A0A226CF51	-----	210
A0A2I7QNH6_(DfrB-H4)	-----KVIGDKYFG--YLDEAKGW--	210
A0A0F6R7A8	-----	210
A0A117UUM5	-----	210
A0A1I4V6W3_(DfrB-H5)	-----ICY--QSMFGQTA-----GWADLPOTHERVKWGCVAEGFLKEVPAS--	210

P00383_(DfrB1)	-----	280
Q4VCS1_(DfrB-H2)	-----	280
P00384_(DfrB2)	-----	280
Q71RY0_(DfrB3)	-----	280
B0FMU1_(DfrB4)	-----	280
A0A0U2TX79_(DfrB5)	-----	280
Q1HVJ4_(DfrB6)	-----	280
D2Y395_(DfrB7)	-----	280
R4UB13_(DfrB9)	-----	280
A0A2N2TNN4_(DfrB11)	-----	280
A0A0A0YVN9	-----	280
F4ZVS8	-----	280
A0A0A0YXA2	-----	280
A0A0A0Z1Q9	-----	280
A0A0A0YWM8	-----	280
A0A0N7AFD7	-----	280
A0A2I5YL54	-----	280
Q53CM2	-----	280
A0A0H5ARK7_(DfrB-H3)	-----TEQQPVSW--	280
A0A0J7Y2N3	-----	280
A0A087N9R8_(DfrB-H6)	SKGLCESLADAAIRALTIPDAPASVVDREITTKPVAVGETAAGVTQAPASVEPVADDVLTCDVCLPPATT	280
A0A2E1B2M0	-----	280
H5WLJ1	-----	280
A0A1R1QGT7	-----	280
A0A1G7DL98	-----	280
A0A226CF51	-----	280
A0A2I7QNH6_(DfrB-H4)	-----VETCDKQLS-----	280
A0A0F6R7A8	-----	280
A0A117UUM5	-----	280
A0A1I4V6W3_(DfrB-H5)	-----QAAPVAGVAWLPKMDAPTGV-----PI-----IAKHKPNARMPLGWGT	280

P00383_(DfrB1)	-----	350
Q4VCS1_(DfrB-H2)	-----	350
P00384_(DfrB2)	-----	350
Q71RY0_(DfrB3)	-----	350
B0FMU1_(DfrB4)	-----	350
A0A0U2TX79_(DfrB5)	-----	350
Q1HVJ4_(DfrB6)	-----	350
D2Y395_(DfrB7)	-----	350
R4UB13_(DfrB9)	-----	350
A0A2N2TNN4_(DfrB11)	-----	350
A0A0A0YVN9	-----	350
F4ZVS8	-----	350
A0A0A0YXA2	-----	350
A0A0A0Z1Q9	-----	350
A0A0A0YWM8	-----	350
A0A0N7AFD7	-----	350
A0A2I5YL54	-----	350
Q53CM2	-----	350
A0A0H5ARK7_(DfrB-H3)	-----QFYQDGKWW	350
A0A0J7Y2N3	-----	350
A0A087N9R8_(DfrB-H6)	VRAGCSFETRLAISAPGRPRHFDEPDTIGSHFGNGGGLDPVRSFHVEDCEASPGFGGDLREAVAVALLW	350
A0A2E1B2M0	-----	350
H5WLJ1	-----	350
A0A1R1QGT7	-----	350
A0A1G7DL98	-----	350
A0A226CF51	-----	350
A0A2I7QNH6_(DfrB-H4)	-----QERIDKGGW	350
A0A0F6R7A8	-----	350
A0A117UUM5	-----AAW	350
A0A1I4V6W3_(DfrB-H5)	IKLG-----HGDEEET-----DALDTI-----LHYNGDSLIIW	350

```

P00383_(DfrB1) -----MERSSNEVSNPVAGNFVFPNSAT----- 420
Q4VCS1_(DfrB-H2) -----ERS----E-MERSSNEVSNPVAGNFVFPNSAT----- 420
P00384_(DfrB2) -----MGQSSDEANAPVAGQFALPLSAT----- 420
Q71RY0_(DfrB3) -----MDQHNGVSTLVAGQFALPSHAT----- 420
B0FMU1_(DfrB4) -----MNEGKNEVSTSAAGRFAFPNSAT----- 420
A0A0U2TX79_(DfrB5) -----MDQGRSEVSNPVAGQFAPNSAA----- 420
Q1HVJ4_(DfrB6) -----MDQGSNEVINPVAGQFASPNSAT----- 420
D2Y395_(DfrB7) -----MDQGSNEVGNPVAGQFSPNSAA----- 420
R4UB13_(DfrB9) -----MNQSSNCISTPVVGGQFALPFQPT----- 420
A0A2N2TNN4_(DfrB11) -----MDQSSKEVGTVPVGGQFALPSHAT----- 420
A0A0A0YVN9 -----MNEGKNEVSTSAAGRFAFPNSAT----- 420
F4ZVS8 -----MNEGKNEVSTSAAGRFAFPNSAT----- 420
A0A0A0YXA2 -----MNEGKNEVSTSAAGRFAFPNSAT----- 420
A0A0A0Z1Q9 -----MNEGKNEVSTSAAGRFAFPNSAT----- 420
A0A0A0YWM8 -----MNEGKNEVSTSAAGRFAFPNSAT----- 420
A0A0N7AFD7 -----ERS----E-MERSSNEVSNPVAGNFVFPNSAT----- 420
A0A2I5YL54 -----MGQSSDEANTPVAGQFALPLGAT----- 420
Q53CM2 -----MGQSSHEANAPVAGQFALPLSAT----- 420
A0A0H5ARK7_(DfrB-H3) N-----GDDRI----KDRKNTAAGIPVRDLYAAPIAQTA----- 420
A0A0J7Y2N3 -----EAERAAPNVAK-GRNAVTFHDQLECEKGKWLGLAALAALPPVKQSV----- 420
A0A087N9R8_(DfrB-H6) KR-----EAERAAPNVAK-GRNAVTFHDQLECEKGKWLGLAALAALPPVKQSV----- 420
A0A2E1B2M0 -----NEATPAAIELNEGLCHSWPATHKT----- 420
H5WLJ1 -----SSFQGNVGGEMVDPWPADAK----- 420
A0A1R1QGT7 -----MVDWPADAK----- 420
A0A1G7DL98 -----MVDWPADAK----- 420
A0A226CF51 -----MVDWPADAA----- 420
A0A2I7QN6_(DfrB-H4) -----LPNLLKELENQMNI-DKVETQTEHQEEMSEFGGEDLPKPLW----- 420
A0A0F6R7A8 -----KRRSGDLVLIDREGMRE----- 420
A0A117UUM5 -----KRRSGDLVLIDREGMRE----- 420
A0A1I4V6W3_(DfrB-H5) NFHGCWEGWVPIGAAEPIAPPPGDPERLDTVYSTEHQRKRLASRVGPADEDLARVVTAMQWAEFVPGQ 420

```

```

P00383_(DfrB1) -----FGMGDRVRKKSгааawqgQIVGWYCTNLTPEGYAVESEAHPGSVQIYPVAALER 490
Q4VCS1_(DfrB-H2) -----FGMGDRVRKKSгааawqgQIVGWYCTNLTPEGYAVESEAHPGSVQIYPVAALER 490
P00384_(DfrB2) -----FGLGDRVRKKSгааawqgQVVGWYCTKLTPEGYAVESESHPGSVQIYPVAALER 490
Q71RY0_(DfrB3) -----FGLGDRVRKKSгааawqgQVVGWYCTKLTPEGYAVESESHPGSVQIYPVAALER 490
B0FMU1_(DfrB4) -----FALGDRVRKKSгааawqgRIVGWYCTTLTPEGYAVESESHPSVQIYPMTALER 490
A0A0U2TX79_(DfrB5) -----FGMGDRVRKKSгааawqgQIVGWYCTKLTPEGYAVESEAHPGSVQIYPVAALER 490
Q1HVJ4_(DfrB6) -----FGMGDRVRKKSгааawqgQIVGWYCTKLTPEGYAVESEAHPGSVQIYPVAALER 490
D2Y395_(DfrB7) -----FSMGDRVRKKSгааawqgQIVGWYCTKLTPEGYAVESEAHPGSVQIYPVAALER 490
R4UB13_(DfrB9) -----FGLGDRVRKKSгааawqgKVVGWYCTKLTPEGYAVESEAHPGSVQIYPVAALER 490
A0A2N2TNN4_(DfrB11) -----FGLGDRVRKKSгааawqgQVVGWYCTKLTPEGYAVESESHPGSVQIYPVNALE 490
A0A0A0YVN9 -----FALGDHVRKKSгааawqgRIVGWYCTTLTPEGYAVESESHPGSVQIYPMTALER 490
F4ZVS8 -----FAWGDRVRKKSгааawqgRIVGWYCTTLTPEGYAVESESHPGSVQIYPMTALER 490
A0A0A0YXA2 -----FALGDRVRKKSгааawqgRIVGWYCTTLTPEGYAVESESHPGSVQIYPMTALER 490
A0A0A0Z1Q9 -----FALGDRVRKKSгааawqgRIVGWYCTTLTPEGYAVESESHPGSVQIYPMTAPER 490
A0A0A0YWM8 -----FALGDRVRKKSгааawqgRIVGWYCTTLTPEGYAVESESHPGSVQIYPMTALER 490
A0A0N7AFD7 -----FGIGDRVRKKSгааawqgQIVGWYCTNLTPEGYAVESEAHPGSVQIYPVAALER 490
A0A2I5YL54 -----FGLGDRVRKKSгааawqgQIVGWYRTKLTPEGYAVESESHPGSVQIYPVAALER 490
Q53CM2 -----FGFGDRVRKKSгааawqgQVVGWYCTKLTPEGYAVESESHPGSVQIYPVAALER 490
A0A0H5ARK7_(DfrB-H3) -----PQGKFRMGDIYKKSSTGSEWEGRVVGYSTEQTKEGYAVESEAHAGSVQIYPKRW 490
A0A0J7Y2N3 -----FNRGDHEKISGSKWRGVVGYEYSTLTPEGYAVESDTETGVSQIYPKALRS 490
A0A087N9R8_(DfrB-H6) -----GKFORGDHVEKVSNSWRGKVVGEYSTDLTPEGYAVESDTETGVSQIYPKALRL 490
A0A2E1B2M0 -----PKFLRGDPVRKRAGSSWHGIIVGEYSTDLTSEGCVESLFEKGSVQIYPAAALE 490
H5WLJ1 -----FRMGDKVRKVRGQWHGRVVGWYSTDLTPEGYAVESDTERGVSQIYPASALE 490
A0A1R1QGT7 -----FQMGDYAAKKGASWRGRIVGWYRTDLTRLGYAIESYFEPGSVQIYPETAIDA 490
A0A1G7DL98 -----FQMGDYAAKKGASWRGRIVGWYRTDLTRLGYAIESYFEPGSVQIYPETAIDA 490
A0A226CF51 -----FQMGDYAAKKGASWRGKIVGWYRTDLTSLGYAIESHFEFEPGSVQIYPETAIEA 490
A0A2I7QN6_(DfrB-H4) -----AKWSVGDSTTKGSSWTGKVVGYSTLTPNGYAVESLTEKGSVQIYPEAALC 490
A0A0F6R7A8 -----GRFGGERVTKTKGSSWTGRVVGfySTELTPIGYAVESETEKGSVQIYPEAALTA 490
A0A117UUM5 -----RKFTLGQRVTKTKGSKWTRVVGfySTNLTVPYVAIESENEPGSVQIYPEAATA 490
A0A1I4V6W3_(DfrB-H5) LVPAFRPDGTGGRDQTGIQLGARVRKTKGSSWQGVVGYATATLPRGVCVIESEREPGSVQIYPAALEP 490

```

* . * : * * : * * * * * * . : * * . * * * * *

```

P00383_(DfrB1) IN----- 560
Q4VCS1_(DfrB-H2) IN----- 560
P00384_(DfrB2) VA----- 560
Q71RY0_(DfrB3) VA----- 560
B0FMU1_(DfrB4) VA----- 560
A0A0U2TX79_(DfrB5) IN----- 560
Q1HVJ4_(DfrB6) VN----- 560
D2Y395_(DfrB7) INGVQG----- 560
R4UB13_(DfrB9) VA----- 560
A0A2N2TNN4_(DfrB11) VA----- 560
A0A0A0YVN9 VA----- 560
F4ZVS8 VA----- 560
A0A0A0YXA2 VA----- 560

```

A0A0A0Z1Q9	VA-----	560
A0A0A0YWM8	VT-----	560
A0A0N7AFD7	IN-----	560
A0A2I5YL54	VAQQFAP-----	560
Q53CM2	VA-----	560
A0A0H5ARK7_(DfrB-H3)	WMTSADTRLTP-----	560
A0A0J7Y2N3	TSAAALHOGSG-----	560
A0A087N9R8_(DfrB-H6)	SHDQRGGSGE-----	560
A0A2E1B2M0	ITTAKPPTRQEFIDRFVKK-MVEVAGE---RFTDGHSI-----ADYAKEIAP--TFDDPS	560
H5WLJ1	EA-----	560
A0A1R1QGT7	WQPPALERE-----	560
A0A1G7DL98	WQPPALERE-----	560
A0A226CF51	WVPPKAESELE-----	560
A0A2I7QNH6_(DfrB-H4)	IETPQQREDRE-----RLEAAYELYCHVIDKETT-----FDKCFCTFGPLKAMYIK--	560
A0A0F6R7A8	APKPELE-----	560
A0A117UUM5	ILGEGK-----	560
A0A114V6W3_(DfrB-H5)	VEAETPTHPTSPQGGEVKAPQDPVGEAK--ALLDKESICRNPWAQVKWLRLLVADQYAATTSEQRAFLDRA	560
P00383_(DfrB1)	-----	586
Q4VCS1_(DfrB-H2)	-----	586
P00384_(DfrB2)	-----	586
Q71RY0_(DfrB3)	-----	586
B0FMU1_(DfrB4)	-----	586
A0A0U2TX79_(DfrB5)	-----	586
Q1HVJ4_(DfrB6)	-----	586
D2Y395_(DfrB7)	-----	586
R4UB13_(DfrB9)	-----	586
A0A2N2TNN4_(DfrB11)	-----	586
A0A0A0YVN9	-----	586
F4ZVS8	-----	586
A0A0A0YXA2	-----	586
A0A0A0Z1Q9	-----	586
A0A0A0YWM8	-----	586
A0A0N7AFD7	-----	586
A0A2I5YL54	-----	586
Q53CM2	-----	586
A0A0H5ARK7_(DfrB-H3)	-----	586
A0A0J7Y2N3	-----Q	586
A0A087N9R8_(DfrB-H6)	-----	586
A0A2E1B2M0	QRVEGPESCALADIDCWE-----R	586
H5WLJ1	-----	586
A0A1R1QGT7	-----	586
A0A1G7DL98	-----	586
A0A226CF51	-----	586
A0A2I7QNH6_(DfrB-H4)	-----IVDKTNYRKGV----K	586
A0A0F6R7A8	-----	586
A0A117UUM5	-----	586
A0A114V6W3_(DfrB-H5)	DAIEA----ALATIRSSQAGEVRDDA	586

Figure S4.5. Sequence alignment of all 30 sequences identified and presented in Figure 4.2.
The alignment was performed with MAFFT.

	DfrB1	DfrB-H2	DfrB-H3	Dfr-BH4	DfrB-H5	DfrB-H6
DfrB1		100	63.5	71.4	62.7	74.0
DfrB-H2	79.6		61.4	50.5	57.4	74.5
DfrB-H3	30.5	35.7		44.4	66.0	56.4
DfrB-H4	20.6	23.0	30.3		49.6	47.0
DfrB-H5	14.0	15.9	20.4	28.6		37.8
DfrB-H6	10.0	11.9	13.5	16.6	18.8	

Figure S4.6. Sequence similarity (%) of DfrB-H.

The upper section is Smith-Waterman local alignment using EMBOSS water², whereas the lower section is Needleman-Wunsch global alignment using EMBOSS needle².

	DfrB1	DfrB-H2-Seg	DfrB-H3-Seg	DfrB-H4-Seg	DfrB-H5-Seg	DfrB-H6-Seg
DfrB1		98.3	55.2	55.2	56.9	55.9
DfrB-H2-Seg	100		55.2	55.2	56.9	55.9
DfrB-H3-Seg	67.2	67.2		46.8	44.8	60.3
DfrB-H4-Seg	69.0	69.0	56.5		60.3	62.1
DfrB-H5-Seg	69.0	69.0	56.9	67.2		52.5
DfrB-H6-Seg	69.5	69.5	70.7	65.5	59.3	

Figure S4.7. SH3-like sequence matrix.

The upper section is the sequence identity (%) using Needleman-Wunsch global alignment with EMBOSS needle², whereas the lower section represents sequence similarity (%) using the same algorithm. Residues 21 to 78 are used for all segments.

DfrB1	-----	0
DfrB-H2	-----	0
DfrB-H3	-----	0
DfrB-H4	-----	0
DfrB-H5	-----	0
DfrB-H6	MLVRTRDIDCCSCGSRPRHRIFLYRAETERIERSIRASPGGKAFHRSIVAASEGLTMT	60
DfrB1	-----	0
DfrB-H2	-----	0
DfrB-H3	-----	0
DfrB-H4	-----	0
DfrB-H5	-----	0
DfrB-H6	NDKDGAALLREAIHNALSSELGDTYDCGRVWDAAWHVGTMGEDDFTPVNDRVGEITETVLA	120
DfrB1	-----	0
DfrB-H2	-----	0
DfrB-H3	-----	0
DfrB-H4	-----	0
DfrB-H5	-----MTEHHTPDAAS-----GTRDG-----	16
DfrB-H6	ALQSHGQAFDAASPLGDFQCHNCKRIVERLTVAPEKCPDCMCSFSAVYGSSEAKYRAAPQ	180
DfrB1	-----	0
DfrB-H2	-----	0
DfrB-H3	-----	0
DfrB-H4	-----	0
DfrB-H5	-----EK PANIQPGELLWICYQSMFGQTA-----WGWADLP THERVKWGCVAE	59
DfrB-H6	YAFDAAGVREAVADAIWAEFMKMAPGYLLSKGLCESLADAAIRALTIPDAP----ASVVD	236
DfrB1	-----	0
DfrB-H2	-----	0
DfrB-H3	-----MSEIKRY	7
DfrB-H4	-----MK-----YTAEQIKS	10
DfrB-H5	GFLKEVPASQAAPVAGVAWLPMDAPTGVPIAKHKPNARMPLGWGTIKLGHGDE----	115
DfrB-H6	REITTKPVAVGETAAGVTQAPA-----SVEPVADDVLTCDVKLPPATTVRAGCSFETLRL	291
DfrB1	-----	0
DfrB-H2	-----	0
DfrB-H3	EPINIDGTCGICAEDADGAYDYRDFEAL-----NDELAALKDGGAPAQ	50
DfrB-H4	ILENAPSGSQ---FYTDDYDVYKRVIGDKYF---GYLDEAKG---WVETCDKQL---	56
DfrB-H5	-----EETDAL-----DTI-LHYNGDSL IWNFHGCWEGWPIGAAEPIAPPP	156
DfrB-H6	-AISAPGRPR---HFDEP-----DTIGSHFGNGGLDPVRSF---HVEDCEASPGFGGG	338
DfrB1	-----MERSSN---E-----V--SNPV-	12
DfrB-H2	-----MQRVVGPHRTPRSSQERSEMERSN---E-----V--SNPV-	31
DfrB-H3	TEQQPVSQFYQDGKWWNGDDRI-----KDHKR-NTE-----A-AGIPVR	88
DfrB-H4	-SQE-----RIDKGWWL PNLKEL ENQMNI DKVETQTEHQE---E-----M--SEFG-	97
DfrB-H5	GDPE-----RLDLTVYSTEHRKRLASRVGPAEDL---ARVVTAMQWAEFVPGQLV-	206
DfrB-H6	DLRE-----AVALW KREAERAAPN--VAKGRNAVTFHDQLECEK GKWLGLADAALA-	389
SH3-like domain		
DfrB1	----AGNFVFPSDATFGMGDRVRKKS GAAWQGGQIVGWYCTNLTPEGYAVESEAHPGSVQI	68
DfrB-H2	----AGNFVFP SNATFGMGDRVRKKS GAAWQGGQIVGWYCTNLTPEGYAVESEAHPGSVQI	87
DfrB-H3	DLYAAPIAQTAPQGKFRMGDIVKKTSGSEWEGRVVGYSTEQTKEGYAVESEAHAGSVQI	148
DfrB-H4	----GEDLPKPLWAKWSVGD SVTKTGSSWTGKVVGYSTSLTPNGYAVESL TEKGSVQI	153
DfrB-H5	PAFRPDGTGGRDQTGIQLGARVRKTGSSWQGVVGYATALT PRGVCVESEREPGSVQI	266
DfrB-H6	----ALPPVKQSVGK FQRGDHVEK VSGSNWRGKVVEGYSTDLTPEGYAVESDTETGSVQI	445
* * * . * : * * : * * * * . * . * * * . * * * * *		
B1 B2 B3 B4		
DfrB1	<u>YPVAALERIN</u> -----	78
DfrB-H2	YPVAALERIN-----	97
DfrB-H3	YPAKRWRQWMTSADTRLTP-----	167
DfrB-H4	YPEAALCDIETPQQ-----REDRERLEAAYELYCHVIDKETTDFDKFCTFGPLKAMYI	205
DfrB-H5	YPIAAL EPVEAETPHTPTSPQGGEVKAPQDPVGEAKALLDKES----ICRN PWAQVKWL	322
DfrB-H6	YPAKALRLSHDQRG-----GSGE-----	463
**		
B4 B5		
DfrB1	-----	78
DfrB-H2	-----	97
DfrB-H3	-----	167
DfrB-H4	KIVDKTNYRKGVK-----	218
DfrB-H5	RLVADQ--YAATTSEQR AFLDRADAIEAALATIRSSQAGEVRDDA	365
DfrB-H6	-----	463

Figure S4.8. Sequence alignment of DfrB1 and the DfrB-H.

The beta strands of the SH3-like fold of DfrB1 are represented; the active-site residues are underlined and in bold. The alignment was performed with Clustal Omega.

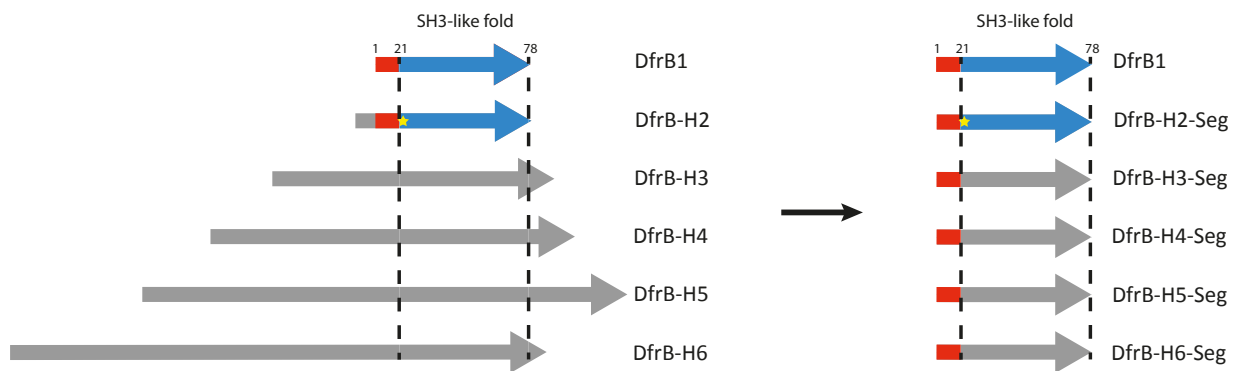


Figure S4.9. Representation of the generation of DfrB-H-Seg sequences from their respective DfrB-H sequence.

The first 20 amino acids of DfrB1 are necessary for its expression. To ensure the expression of each DfrB-H-Seg, these 20 amino acids were fused to the N-terminus of the SH3-like fold of each DfrB-H. DfrB-H sequences are schematically aligned to DfrB1. Sequence lengths are not to scale. The SH3-like sequence of DfrB1 is colored in blue. The yellow star represents a substitution in the sequence. The N-terminus of DfrB1 is colored in red.

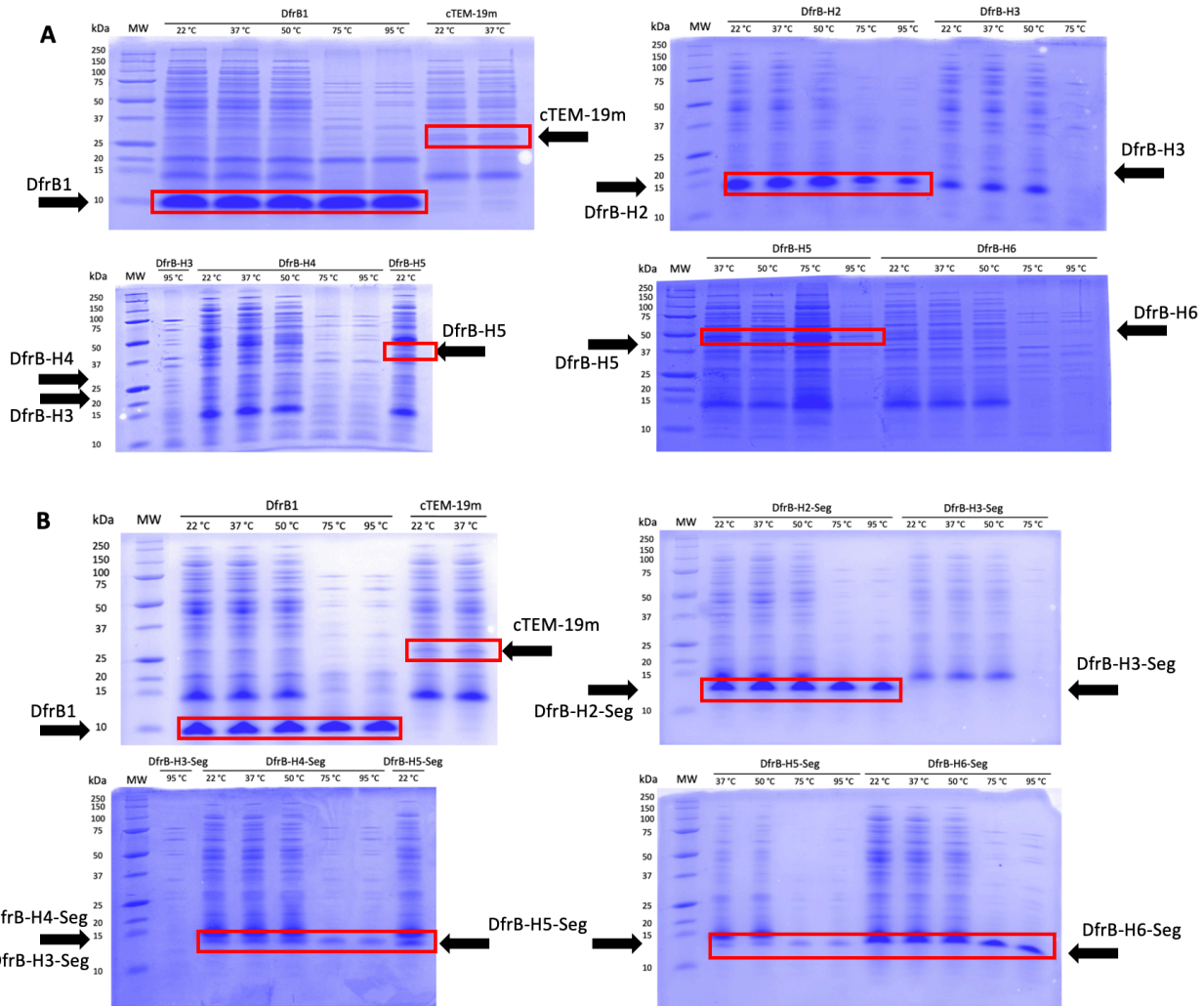


Figure S4.10. Tricine-SDS-PAGE of the heated lysate of overexpressed proteins for which dihydrofolate activity has been determined.

Temperatures at which each lysate was incubated for 10 minutes, before cooling on ice and testing for activity, are indicated. **A.** Lysate obtained following overnight expression of cells at 37°C. **B.** Lysate obtained following expression of cells for 3h at 37°C. Molecular weights: DfrB1: 11.0 kDa; cTEM-19m: 30.8 kDa; DfrB-H2: 12.8 kDa; DfrB-H3: 20.9 kDa; DfrB-H4: 27.3 kDa; DfrB-H5: 41.7 kDa; DfrB-H6: 51.7 kDa; DfrB-H2-Seg: 11.0 kDa; DfrB-H3-Seg: 11.4 kDa; DfrB-H4-Seg: 10.9kDa; DfrB-H5-Seg: 10.9 kDa; DfrB-H6-Seg: 11.0 kDa.

We note that overexpression of DfrB-H5-Seg in *E. coli* without antibiotic selection yielded low cell viability, suggesting dosage toxicity. Toxicity has been observed upon overexpression of dihydrofolate reductases³. Although the mechanism leading to DfrB-H5-Seg toxicity is unknown, it is gene-specific since this toxicity was not observed for the DfrB or other DfrB-H-Seg.

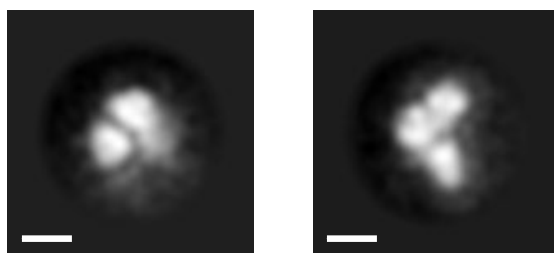


Figure S4.11. DfrB-H5 assembles into various multimeric assemblies.

In addition to tetrameric species (main manuscript), negative stain electron microscopy images of DfrB-H5 reveals the assembly of dimeric (left) and trimeric (right) species, assembled respectively with 299 and 253 particles. The white bar corresponds to 50 Å

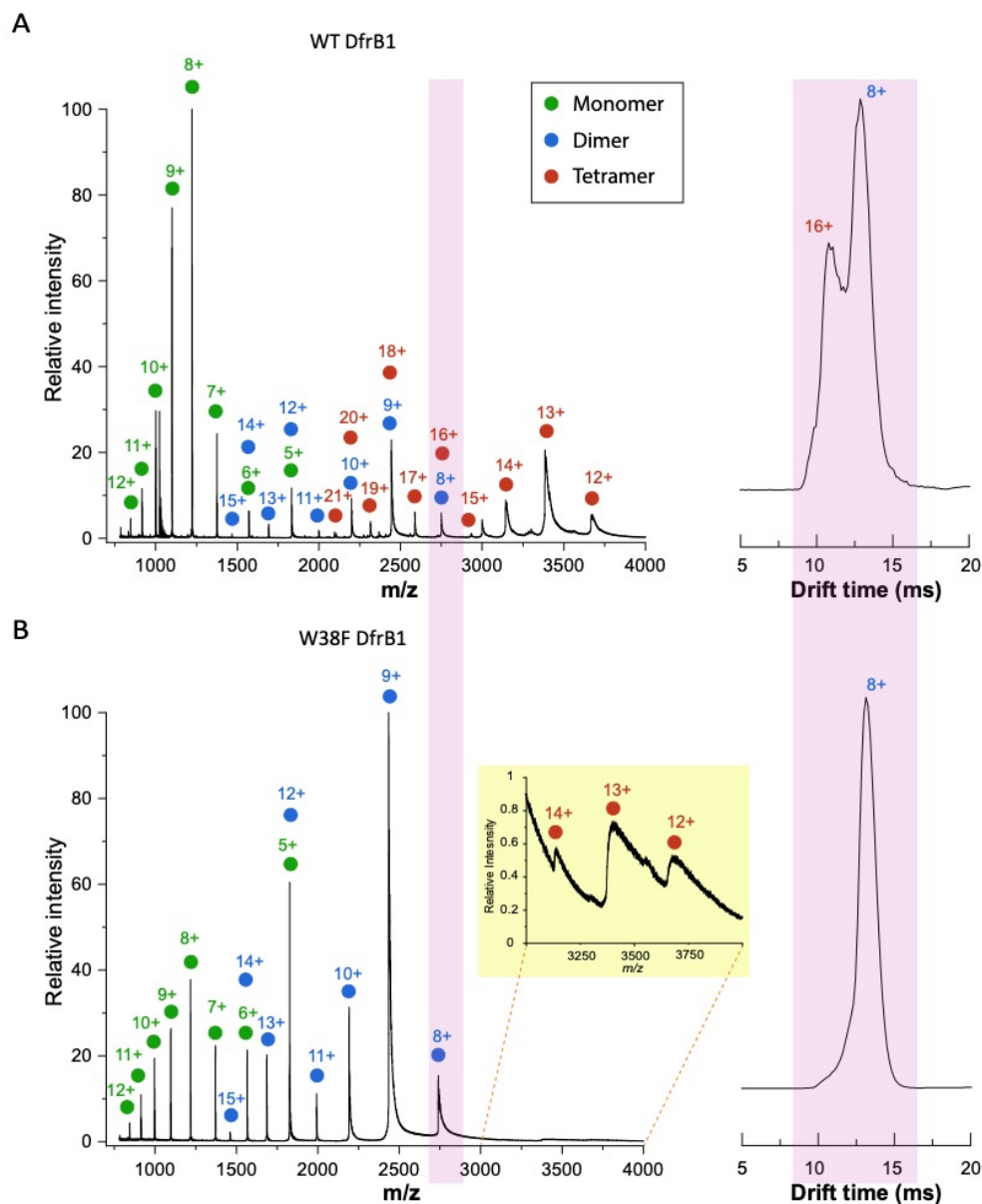


Figure S4.12. Native nanoESI-MS and ion mobility studies demonstrate that the W38F substitution in DfrB1 nearly abolishes the tetramerisation observed in WT DfrB1.

A. Native mass spectrum of WT DfrB1 (left) and ion mobility arrival time distribution (right) for the peak highlighted in pink ($m/z = 2749.64$). The peaks are annotated by their ionic charge and are colored in green (monomer), blue (dimer) or red (tetramer) according to their oligomeric state. Certain oligomeric charge states gave overlapping signals in m/z space, but these isobaric ions could be easily resolved by ion mobility (right panels), which separates ions on the basis of charge and shape. **B.** Native mass spectrum of W38F DfrB1 (left) and ion mobility arrival time distribution (right) of the peak highlighted in pink ($m/z = 2739.95$). The inset zoom shows that the tetramer of W38F DfrB1 is present in trace quantities. The drift time distribution of the peak highlighted in pink includes both dimeric (8+) and tetrameric (16+) states for WT DfrB1, whereas W38F DfrB1 shows the presence of only the dimeric (8+) state.

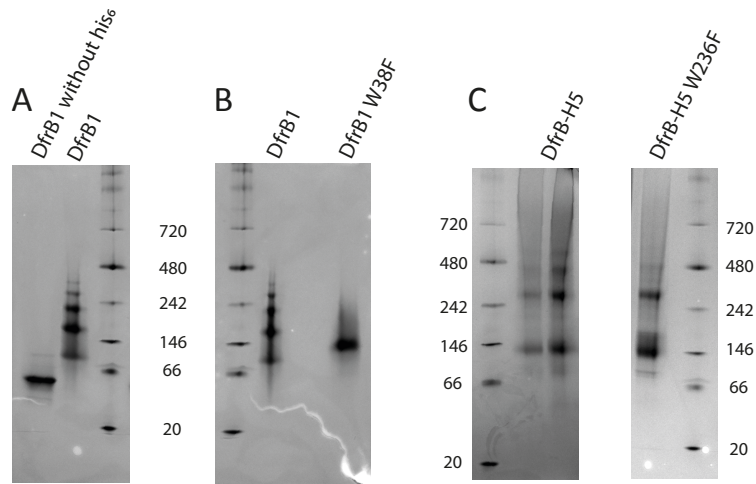


Figure S4.13. Oligomerisation observed with native PAGE.

A. Protein migration on native PAGE is influenced by the presence of the His₆-tag on DfrB1. DfrB1 runs exclusively as a tetramer when no His₆-tag is present. The addition of 24 histidine residues (His₆ fusion tag for each of four monomers) to the low molecular weight DfrB1 (78 residues per monomer) has an important impact on its migration in native PAGE, increasing the number of forms observed. We note that the kinetic and inhibition properties of DfrB1 are only mildly modified by the presence of the His₆ tag ⁴. **B.** The W38F mutation in DfrB1 modifies the oligomerisation state as observed on gel. **C.** The W236F mutation in DfrB-H5 does not modify the oligomerisation state as observed on gel. The difficulty in obtaining sufficient quantities of His₆-DfrB-H5 unfortunately precluded its analysis in the absence of a His₆-tag. DfrB1 without His₆ tag: 8.4 kDa, DfrB1: 11.0 kDa, DfrB-H5: 41.7 kDa.

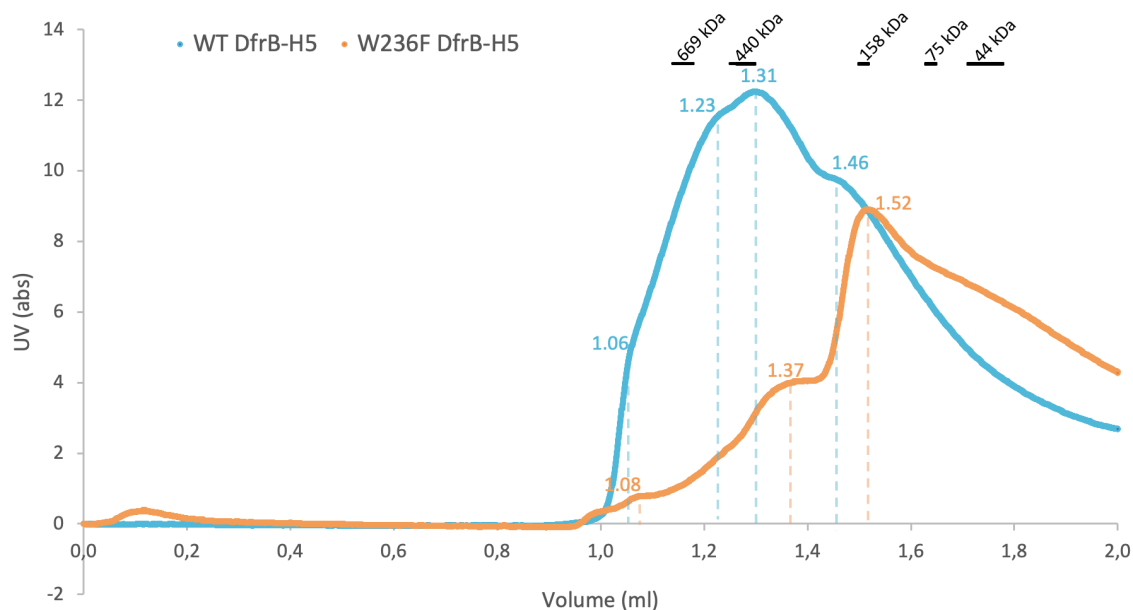


Figure S4.14. SEC chromatograms of WT and W236F DfrB-H5.

2.5 mg/mL DfrB-H5 WT and DfrB-H5 W236F. Both proteins present several peaks corresponding to the multimerisation states. The monomeric form is expected to form a peak at 1.64 mL. Separation was performed on a 2.4 mL Superdex 200 Increase 3.2/300. The molecular weight of five standards are indicated above the graph according to their retention time.

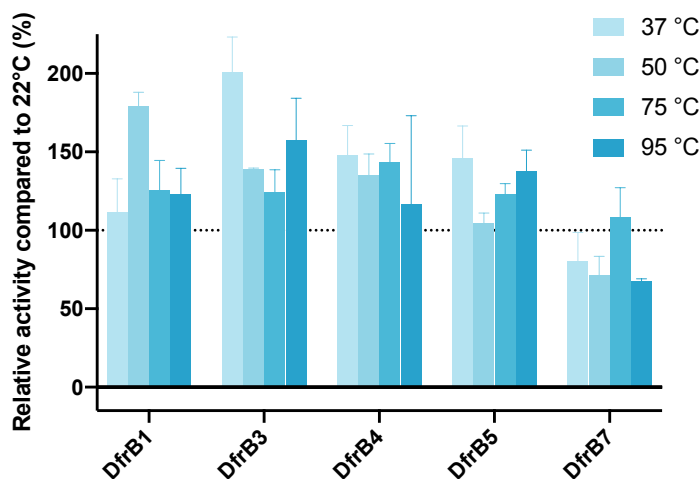


Figure S4.15. The members of the DfrB family are tolerant to heating.

Dihydrofolate reductase activity was assayed in *E. coli* lysates following heating at various temperatures (from 37°C to 95°C, as indicated), followed by cooling and centrifugation. It is not possible to assay dihydrofolate reductase activity at high temperature due to the lability of the DHF substrate and NADPH cofactor⁵. Values greater than 100% should be interpreted with caution, as they may result from a reduction in volume of the supernatant following the process of heat-induced precipitation.

4.12 Supplementary references

- (1) Krahn, J. M.; Jackson, M. R.; DeRose, E. F.; Howell, E. E.; London, R. E. Crystal Structure of a Type II Dihydrofolate Reductase Catalytic Ternary Complex [†]. *Biochemistry* 2007, 46 (51), 14878–14888. <https://doi.org/10.1021/bi701532r>.
- (2) Madeira, F.; Pearce, M.; Tivey, A. R. N.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R. Search and Sequence Analysis Tools Services from EMBL-EBI in 2022. *Nucleic Acids Res.* 2022, gkac240. <https://doi.org/10.1093/nar/gkac240>.
- (3) Bhattacharyya, S.; Bershtein, S.; Yan, J.; Argun, T.; Gilson, A. I.; Trauger, S. A.; Shakhnovich, E. I. Transient Protein-Protein Interactions Perturb E. Coli Metabolome and Cause Gene Dosage Toxicity. *Elife* 2016, 5, e20309. <https://doi.org/10.7554/eLife.20309>.
- (4) Toulouse, J. L.; Shi, G.; Lemay-St-Denis, C.; Ebert, M. C. C. J. C.; Deon, D.; Gagnon, M.; Ruediger, E.; Saint-Jacques, K.; Forge, D.; Vanden Eynde, J. J.; Marinier, A.; Ji, X.; Pelletier, J. N. Dual-Target Inhibitors of the Folate Pathway Inhibit Intrinsically Trimethoprim-Resistant DfrB Dihydrofolate Reductases. *ACS Med. Chem. Lett.* 2020, acsmedchemlett.0c00393. <https://doi.org/10.1021/acsmedchemlett.0c00393>.
- (5) Ebert, M. C. C. J. C.; Morley, K. L.; Volpato, J. P.; Schmitzer, A. R.; Pelletier, J. N. Asymmetric Mutations in the Tetrameric R67 Dihydrofolate Reductase Reveal High Tolerance to Active-Site Substitutions: Asymmetric Mutations in R67 Dihydrofolate Reductase. *Protein Science* 2015, 24 (4), 495–507. <https://doi.org/10.1002/pro.2602>.

Chapitre 5. L'exploration de l'évolution moléculaire du domaine DfrB

Préface

L'exploration de l'espace des séquences dans le Chapitre 4 nous a renseignés sur la capacité des homologues des DfrB à catalyser la même réaction chimique, avec la même efficacité, que les protéines identifiées dans les échantillons cliniques et associées à la résistance au triméthoprime. Curieusement, même les homologues les plus éloignés sur le plan évolutif que nous avons caractérisés, parmi les trente identifiés dans le Chapitre 4, étaient fonctionnels en tant que dihydrofolate réductases.

Dans le présent chapitre, nous avons poussé plus loin notre exploration de l'espace des séquences du domaine DfrB, en générant nos propres profils de modèles de Markov cachés et en explorant diverses bases de données génomiques et métagénomiques. Cela nous a permis d'identifier des homologues fonctionnels plus éloignés que ceux identifiés dans le Chapitre 4, ainsi que des homologues incapables de catalyser la réduction du dihydrofolate. Nous avons annoté cet espace de séquence avec plusieurs propriétés répertoriées chez la DfrB1, telles que la formation d'un homotétramère, la résistance au triméthoprime et la thermostabilité. Cette analyse de l'espace de séquence, intégrant des informations expérimentales et computationnelles, nous permet de proposer, pour la première fois, un modèle évolutif pour l'émergence de l'activité catalytique dans le domaine DfrB.

Ce chapitre est un manuscrit en préparation. J'ai contribué à la conceptualisation de ce projet, en collaboration avec Prof. Joelle N. Pelletier, Prof. Rachel Kolodny et Prof. Nir Ben-Tal. J'ai identifié les séquences dans les bases de données UniRef, guidée par Prof. Rachel Kolodny. Keigo Ide a effectué la recherche de séquences dans la base de données privée bitBiome, à laquelle le Dr Soichiro Tsuda nous a donné accès, et Dre Janine N. Copp a effectué une recherche dans la base de données métagénomiques JGI/IMG. J'ai généré notre ensemble final de séquences, que j'ai représenté sous la forme d'un arbre phylogénétique et d'un réseau de séquences. J'ai généré les résultats des tests de résistance au triméthoprime. J'ai purifié les homologues et caractérisé leur capacité cinétique et leur multimérisation. Maxime St-Aubin a caractérisé la thermostabilité des homologues, et a généré les données pour caractériser l'affinité des homologues pour le NADPH, que j'ai analysées avec lui. J'ai généré les prédictions d'AlphaFold-multimer à l'aide des ressources de l'Université de Haïfa, guidée par Prof. Rachel Kolodny. J'ai analysé les surfaces électrostatiques. J'ai annoté les génomes des homologues identifiés, et Stella Cellier-Goetghebeur a analysé

ces contextes génomiques et les voies métaboliques associées aux gènes s'y retrouvant. J'ai rédigé la première version de ce manuscrit et généré toutes les figures. Stella Cellier-Goetghebeur et Prof. Joelle N. Pelletier ont contribué à sa modification. Ce manuscrit a été approuvé par tous les auteurs.

Article de recherche 3. A walk through sequence space: identifying the essential features for the emergence of catalytic activity in an SH3 fold

Claudèle Lemay-St-Denis^{1,2,3}, Stella Cellier-Goetghebeur^{1,2,3}, Maxime St-Aubin^{1,2,3}, Keigo Ide⁴, Janine N. Copp⁵, Soichiro Tsuda⁴, Nir Ben-Tal⁶, Rachel Kolodny⁷ and Joelle N. Pelletier^{1,2,3,8}

¹ PROTEO, The Québec Network for Research on Protein, Function, Engineering and Applications, Québec, Canada

² CGCC, Center in Green Chemistry and Catalysis, Montréal, Canada

³ Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, Canada

⁴ bitBiome, Japan

⁵ Michael Smith Laboratories, University of British Columbia, Vancouver, Canada

⁶ Department of Biochemistry and Molecular Biology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

⁷ Department of Computer Science, University of Haifa, Haifa, Israel

⁸ Chemistry Department, Université de Montréal, Montréal, Canada

*Correspondence: joelle.pelletier@umontreal.ca

5.1 Abstract

Enzymes have been pivotal in the evolution of life, demonstrating their remarkable ability to adapt and diversify across billions of years. While extensive research has recapitulated how the catalytic activity of existing enzymes diversifies throughout evolution, the initial emergence of enzymatic activity from non-catalytic proteins remains poorly understood. This study focuses on the Src Homology 3 (SH3) fold, a widely recognized binding module, and specifically examines the only SH3 domain known to be capable of catalysis by itself: the DfrB domain. Unlike typical enzymes, the DfrB domain tetramerizes, giving rise to a unique and symmetrical active site formed by the interface of four identical protomers. How this assembly evolved to be catalytically active is yet unknown. Through a comprehensive exploration of the DfrB sequence space, integrating experimental and computational analyses, we identified an intricate relationship between the formation of the homotetrameric configuration and catalytic activity during the evolution of this domain. Whereas no active site residues are conserved among all catalytically active homologues, DfrB1-like tetramer formation proves to be a powerful predictor of activity, validating the accuracy of AlphaFold-multimer on a large scale for the first time. We propose a model for the evolution of the DfrB domain and demonstrate that electrostatic potential at the central pore emerges as a conserved feature among catalytically competent DfrB homologues. These findings challenge established paradigms on the emergence of catalysis and contribute to the understanding of how enzymatic function evolves from non-catalytic precursors, shedding light on the complex interrelationships that govern protein evolution.

5.2 Introduction

Enzymes are central to life as we know it. Their capacity to evolve novel reactivities in response to their environment has been a fundamental aspect contributing to the diversification of life over the past four billion years. In recent decades, evolutionary biochemists have begun to unravel how catalytic activity within an existing enzyme can evolve into a novel reactivity. The recent evolution of enzymes in response to anthropogenic molecules has provided many valuable examples.¹⁻³ Directed evolution has also been central to this understanding: variants evolved in the laboratory to catalyze non-native chemistry from an existing – often weakly functional or non-specialized – enzyme have been studied and rationalized.^{4,5} Indeed, our growing understanding of the mechanisms of active-site modification has allowed for the rational introduction of new-to-nature reactivities into existing enzymes.⁶⁻⁸

However, fundamental questions, such as how enzymatic catalysis emerges from a non-catalytic protein, are still poorly understood. There are few documented examples of a binding property being up-cycled into a catalytic property; in such cases, the ligand becomes the substrate.⁹⁻¹³ These known cases suggest that the mechanism of emergence includes orienting the substrates for efficient transition state positioning, tuning

of productive and unproductive protein motions and amino acid substitutions in the binding pocket for implementing chemistry conducive to catalysis.¹⁴

The Src Homology 3 (SH3) fold is one of the most ancient folds, and is comprised of a β -barrel formed by five antiparallel β -strands connected through four loops and one of the loops contains a 3_{10} helix.¹⁵ In eucaryotes, this fold is found in multi-domain signaling proteins and has been largely described for its ability to bind peptides with proline-rich motifs. Its role in protein-protein interactions is key for the assembly of complexes.¹⁶ In these cases, the loops between the B1-B2 strands and B2-B3 strands are important determinants of the binding specificity.¹⁶ The function of prokaryotic SH3-fold domains are more diverse, as they were reported to bind peptides, proteins, DNA, RNA as well as metals.^{16,17} While the vast majority of SH3-fold domains do not form homocomplexes, a few have been reported, but no common surface is exploited in these homomers.¹⁸⁻²⁰ At the moment, there are 16,000 SH3 domain entries in the ECOD database, representing 1.6% of this protein domain database.²¹

Fascinatingly, only one evolutionary iteration of the SH3 fold has been reported to be sufficient and essential to catalyze a chemical reaction: the DfrB domain. The type B dihydrofolate reductase (DfrB) protein family forms its unique and symmetrical central active site through the interface of four identical SH3 folds (Figure 5.1A).²² The substrate dihydrofolate (DHF) and the reducing cofactor NADPH bind in this central tunnel; a hydride is transferred from the nicotinamide group of NADPH to the pterin group of DHF (Figure S5.1), making it a dihydrofolate reductase.

Notably, the DfrB proteins lack the typical features of enzymes. The homotetrameric DfrB structure results in a symmetry in its active site, where the four identical B4 strands line the tunnel, contributing the same residues to establish different interactions between the substrate and the cofactor (Figure 5.1A).²² Consequently, optimizing binding by engineering these residues presents a complex problem, as establishment of new interactions specific to one ligand will generally be disruptive to binding the other ligand. Furthermore, active site engineering has shown some instances where simultaneous mutations of all four active site residues on the B4 strand procure new active site environments that include up to 16 substitutions (4 substitutions \times 4 B4 strands), with maintenance or even enhancement of activity.²³ This is explained by the fact that the interactions of the protein with the ligands are mainly mediated through interactions with the backbone; in DfrB, catalysis is proximity-based and substrate disorder-assisted, and does not involve any essential chemistry of the enzyme.^{22,24,25} Altogether, this is an unusual approach to enzymatic catalysis. It has the benefit of expanding the binding surface of the small SH3 domain via tetramerization and the drawbacks of high symmetry that reduces binding specificity and of reliance on the non-reactive portion of the substrate to provide catalytically-essential dynamics.

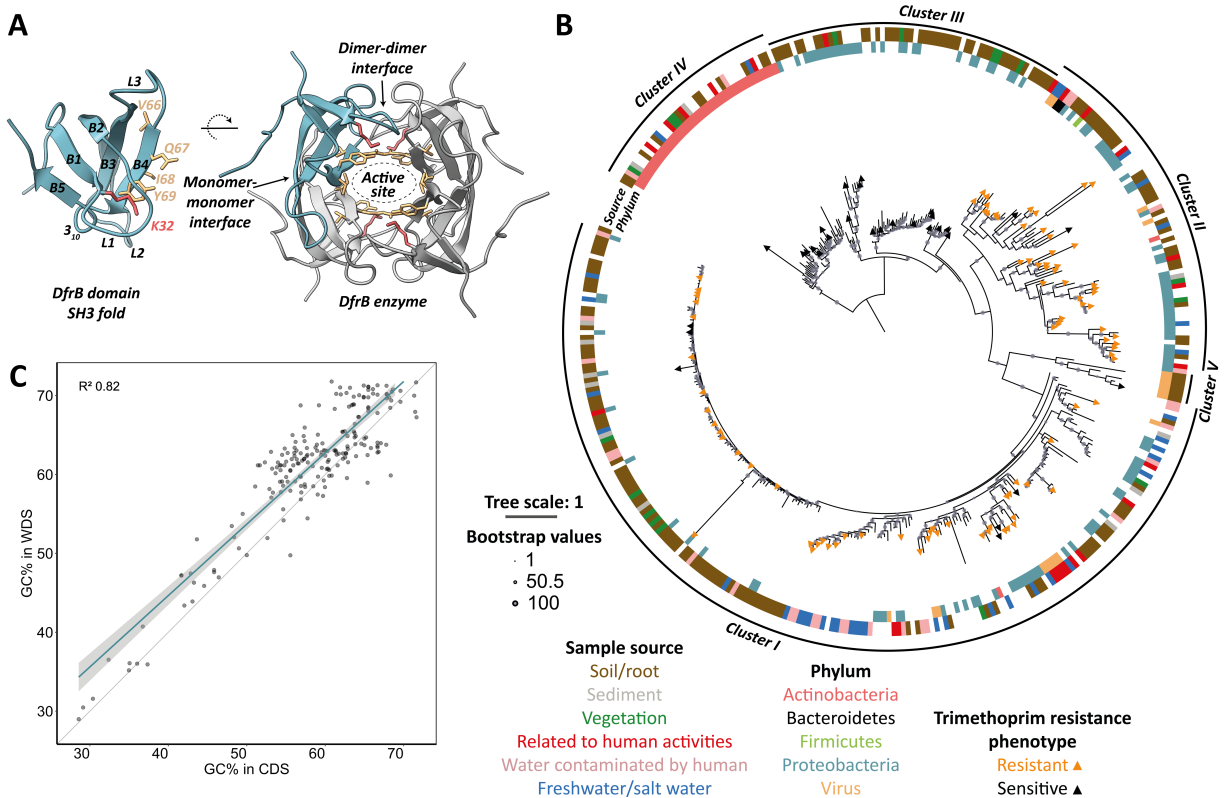


Figure 5.1. Proteins containing the DfrB domain are endogenous to environmental organisms.

A. The topology of the DfrB1 (pdb 2rk1) domain is presented in cartoon representation and annotated, with the side-chains of important catalytic residues represented in stick conformation. The structure of the homotetrameric DfrB1 enzyme is represented, with only one protomer colored in blue. Important elements for the formation of this enzyme are indicated. **B.** Phylogenetic tree of the DfrB domain, rooted at the intersection of sequences from Actinobacteria and from Proteobacteria. The source of the sample in which the protein was identified and the taxonomic phylum of the organism are indicated. The trimethoprim resistance phenotype of proteins when expressed in bacteria is annotated. Clusters are annotated according to the representation in Figure 2, below. **C.** The GC content of *dfrB* homologues is plotted against the GC content of their cognate genome.

How did this simple but unusual catalyst evolve? Here, we explore the sequence space to uncover the properties that have driven the evolution of this SH3-fold based enzyme. The information gathered from this sequence space, guided by predictions of complex formation, electrostatic potential, catalytic activity and genomic context, allows us to propose a model for the evolution of catalytic activity from an SH3 fold.

5.3 Results

5.3.1 The DfrB domain is widely embedded in environmental organisms

To investigate the evolution of the DfrB domain, we collected 386 representative sequences with sequence similarity to the characterized DfrB family members from genomic and metagenomic databases using Hidden Markov Model (HMM) profiles (see *Materials and Methods*). These homologues are shown as a phylogenetic tree (Figure 5.1B).

The DfrB enzymes were first discovered in the context of antibiotic resistance. The faculty of the DfrB protein family to catalyze DHF reduction may have promoted their modern-day distribution in pathogenic bacteria exposed to trimethoprim, a synthetic antimicrobial that selectively targets the ubiquitous and evolutionarily unrelated Fola bacterial dihydrofolate reductase.²⁶ Indeed, we recently reported the first DfrB genes (*dfrB1* to *dfrB9*) near mobile genetic elements (MGE), genomic structures that are central to the horizontal transfer of antibiotic resistance-related genes between pathogenic organisms, suggesting that their capacity to reduce dihydrofolate and genetic mobility are inter-related.^{27,28} However, genetic mobility makes their evolution difficult to track.

Importantly, while the *dfrB1* to *dfrB9* are frequently observed within MGE and in clinical contexts,^{27,28} more recent work and our current analysis demonstrate that this is not the case for their homologues.²⁹ Of the 139 *dfrB* homologues in our current dataset that have an analyzable genomic context (*Materials and Methods*), only one (0.7%) has an integrase gene nearby and five (3.6%) *dfrB* homologues have a transposase gene nearby. In addition, three (2.2%) *dfrB* homologues are associated with incomplete integrons: *attC* sites are in close proximity. An *attC* site is an integron element that leads to gene excision and allows gene mobility when recognized by an integrase, although no integrase gene is detected nearby in these three cases. The rarity of MGE in these contexts indicates that the vast majority (94.2%) of *dfrB* homologues identified here are in different genomic contexts from the first *dfrB* identified in pathogenic bacteria.

Genomic markers other than MGE can also provide clues to the recent evolution of genes. A gene that is endogenous to its host organism will have a similar guanine-cytosine content (%GC) as the genome in which it is embedded. In the current dataset, the %GC of the *dfrB* homologues is proportional to that of their genome; the ratio $\%GC_{\text{gene}}/\%GC_{\text{genome}}$ varies from 0.83 to 1.14 (Figure 5.1C). This suggests that either the *dfrB* homologues are endogenous to their host organisms, or their integration is sufficiently old for their GC content to have adapted. Notably, our previously reported dataset of *dfrB* found in clinical contexts does not show the same behavior; the %GC of the *dfrB* appears to be unrelated to that of their host organism, supporting the hypothesis that *dfrB* in clinical contexts are recently acquired, and have not yet adapted to the GC content of their pathogenic host (Figure S5.2).²⁷

Overall, 63% (117 out of 186) of the DfrB homologues with an assigned host organism are found in Proteobacteria, of which 80% are in Alpha-proteobacteria (Figure 5.1B). Actinobacteria (23%) and viruses (11%) are the other two most common taxonomic groups. Of the 203 homologues for which the source from which they were isolated was reported, 63% were reported to be from soil-related samples, 15% from water-related samples, and 11% from plant/root-related samples (Figure 5.1B). These data are consistent with recent reports that DfrB genes are abundant in rhizosphere and aquatic environments.²⁸⁻³¹

5.3.2 A global view of the DfrB sequence space

The DfrB1 to DfrB9 family members have many distinctive features. These experimentally-determined properties include their characteristic homotetramerization, their capacity to bind the NADPH dinucleotide, their catalytic reduction of dihydrofolate and their thermotolerance – full catalytic activity is recovered after incubation at 95°C.^{22,32} We explored each of these properties within the broader DfrB sequence space mapped here to guide our understanding of the features that may have driven the evolution of the DfrB enzyme. To this end, we represented the DfrB sequence space by a sequence similarity network (SSN) in which each node is a protein, and nodes with sequence similarity above a defined threshold are connected by an edge. Five distinct clusters can be described, with clinically-relevant DfrB members present in cluster *I* (Figure 5.2).

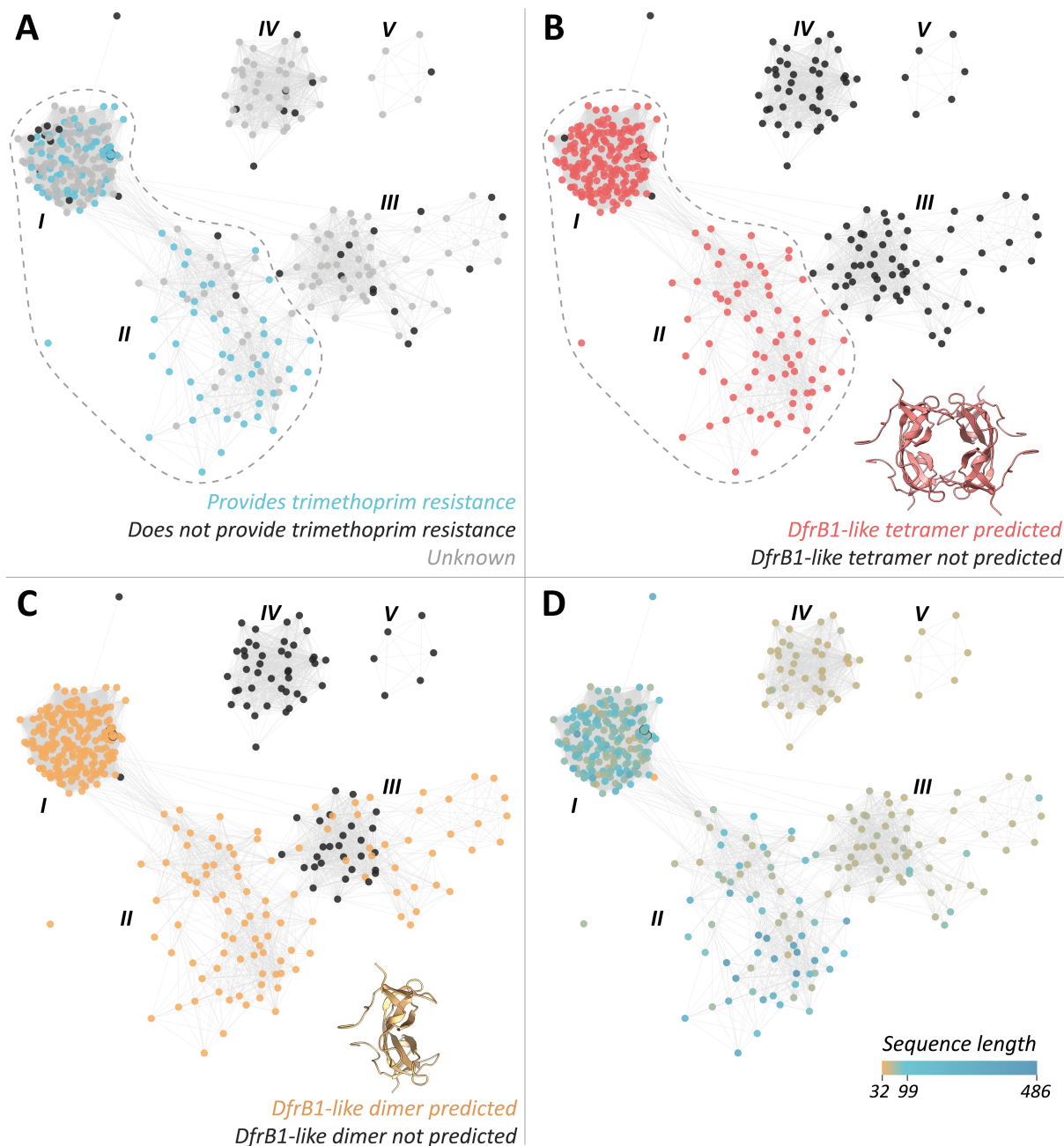


Figure 5.2. Tetramerization is a defining property for the emergence of a catalytic function in the DfrB enzymes.

A. A network visualization of the DfrB sequence space is presented, with each node representing a protein homologue. Edges connect nodes with protein sequences aligning at an E-value lower than 10^{-6} , a metric that quantifies sequence similarity. The DfrB family members identified in a clinical context and retained after the filtration step (DfrB4 and DfrB7) are in cluster *I*, represented by octagons with a black border. Homologues are colour-coded based on their ability to confer resistance to *E. coli* at a minimum concentration of 75 $\mu\text{g}/\text{mL}$ trimethoprim. **B.** Using only the sequence corresponding to the SH3 fold of all homologues, AlphaFold-multimer predicted the formation of tetramers and **C.** of dimers, as described in Figure S5.3. Predicted structures resembling the DfrB complex are coloured. **D.** Homologues are coloured according to their full sequence length. Sequence length spans from 32 to 486 residues, the average length being 99 residues.

To form the DfrB enzyme, two protomers first assemble into a homodimer, followed by dimerization of two homodimers into a homotetramer.²² We predicted the homodimer and homotetramer complexes using only the SH3 fold of each homologue (the region homologous to residues 20 to 78 in DfrB1) with AlphaFold-multimer and compared these predictions to the experimentally-determined DfrB1 complex (pdb 2rk1) (Figure 5.2B,C). To be considered a DfrB1-like tetramer, the monomer-monomer interface must result from contact between two B2 strands of two protomers, and both dimer-dimer interfaces must be formed through contact between the L1 and L3 loops (Figure S5.3).²² The predicted propensity to form a DfrB1-like dimer is shared by the vast majority of homologues in clusters *I* and *II*, whereas it appears to be a transient feature of the homologues in cluster *III*. The predicted propensity to form a DfrB1-like tetramer is more clearly defined in sequence space: a clear demarcation exists between clusters *I* and *II*, where the DfrB1-like tetramer is predicted, and the remaining clusters, where it is not. To our knowledge, no other SH3 domain utilizes the same contacts to form a homocomplex.

High-confidence prediction of the DfrB1-like homotetrameric complex (Figure S5.4), known to be essential for the formation of the central active site of DfrB enzymes, proves to be a robust predictor of catalytic activity in distant homologues. For the 148 homologues tested in the laboratory for their potential to confer trimethoprim resistance when expressed in *E. coli* – a convenient read-out for dihydrofolate reductase activity, since the native Foa dihydrofolate reductase is fully inhibited by trimethoprim – AlphaFold-multimer predictions correlated with DfrB function in 94% of the cases (Figure 5.2A,B). Notably, all homologues where a homotetramer was predicted but its configuration was structurally dissimilar to the DfrB1 complex showed no activity *in vivo*. The exceptions to this relationship were nine homologues that were predicted to form a DfrB1-like tetramer that did not confer trimethoprim resistance. For eight of these, this is thought to be due to their absence of soluble expression, as evidenced by the absence of DfrB homologue overexpression bands on SDS-PAGE (Figure S5.5).

5.3.3 A model of evolution of a catalytically competent active site

Strikingly, the 114 homologues that can catalyze dihydrofolate reduction do not share the same active site residues; any of 18 active site motifs are conducive to catalysis (Figure 5.3A). While VQIY is the most common active site motif, it is exclusive to homologues in cluster *I* which includes all clinically-observed DfrB; homologues in cluster *II* display a wide range of active site motifs. These naturally-evolved sequences, which can catalyze the reduction of dihydrofolate, demonstrate that none of the active-site residues on the B4 strand are essential for catalysis. This supports our previous engineering effort, where we had demonstrated the permissiveness of the active site to patterns of substitution.²³ While it is challenging to definitively identify an overarching trend concerning the identity of catalytically-suitable residues at each position, or patterns of residues, most homologues displaying dihydrofolate reductase

activity share an active site motif characterized by non-polar/polar/non-polar/aromatic residues (Figure 5.3A).

The permissiveness observed in the identity of active site residues within the central and symmetric pore of DfrB homologues suggests that catalysis relies on creating an environment conducive to catalysis rather than requiring specific chemistry within the tunnel. Because the mechanism of DfrB catalysis is proximity-based and substrate disorder-assisted, binding of the substrate and cofactor with catalytically suitable distances and orientations is the principal requirement. In agreement with this, the pore formed by all 18 active site motifs exhibits a positive electrostatic potential (Figure 5.3B, Figure S5.6). This is consistent with their capacity to bind their negatively-charged ligands: the NADPH dinucleotide has phosphate groups and DHF bears carboxylate groups. In contrast, DfrB homologues that are predicted to form a dimer but not a tetramer are not consistently characterized by active site motifs with a positive electrostatic potential.

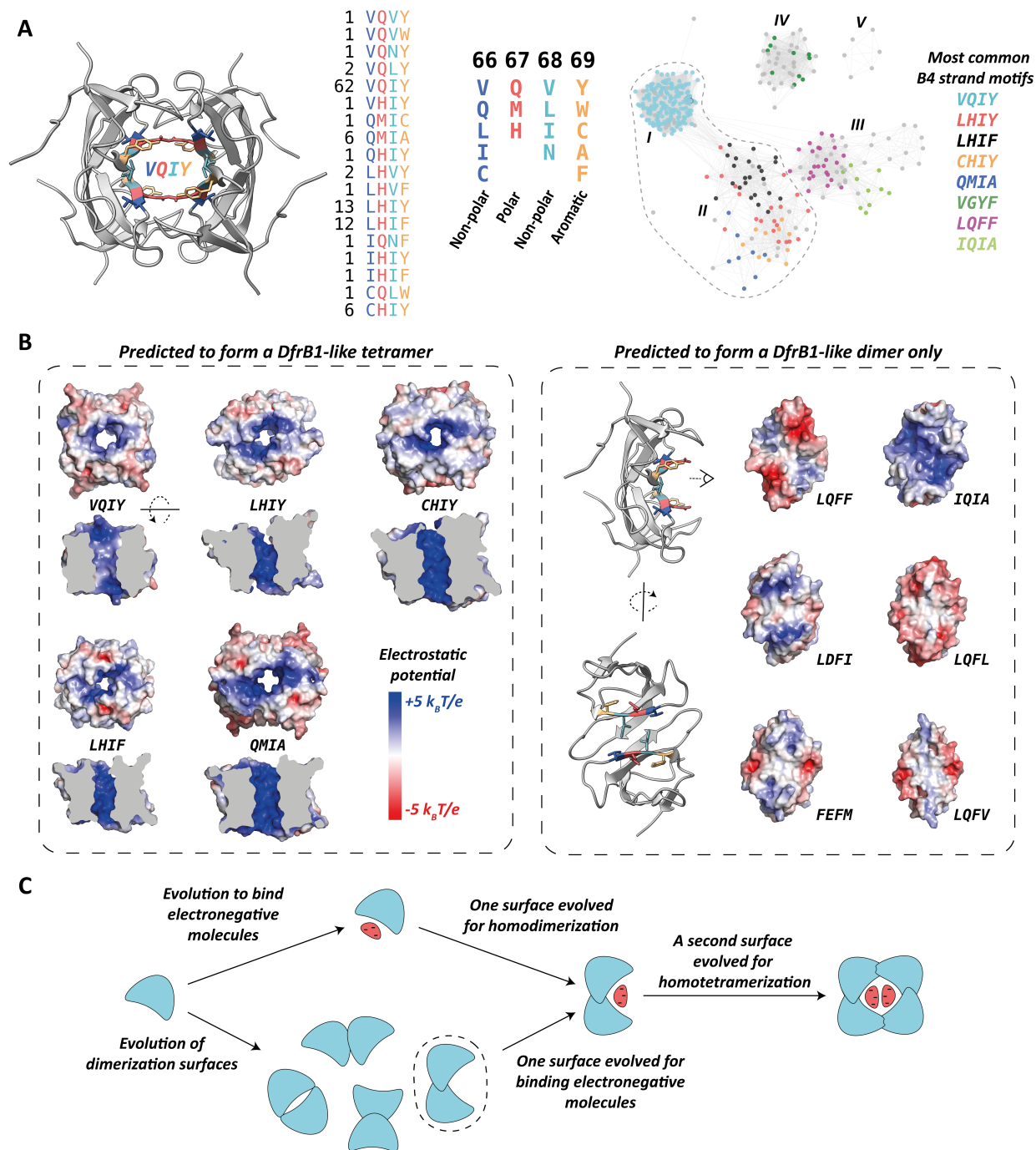


Figure 5.3. The DfrB tunnel is an environment evolved for ligand binding rather than a specialized active site.
A. The B4 strand of the DfrB domain is represented on each protomer, forming the active-site tunnel. All 18 active-site motifs that yield an active dihydrofolate reductase are presented; they correspond to residues 66 to 69 in DfrB1. The number at the left of each motif corresponds to the number of functional homologues having this motif. Homologues displaying one of the eight most common B4 strand motifs are colored on the SSN. **B.** Left: The surface of predicted DfrB1-like tetramers of homologues representing five of the 18 active site motifs is colored according to their electrostatic potential. The 13 remaining active site motifs are similarly represented in Figure S5.6. Right: The surface of predicted DfrB1-like dimers of six representative homologues not predicted to form a DfrB1-like tetramer, is colored according to their electrostatic potential. **C.** A model of possible pathways for the evolution of the DfrB domain capable of catalyzing the reduction of dihydrofolate is proposed.

We further demonstrated that catalysis in the DfrB enzyme depends on oligomerization rather than the active site chemistry by in-depth characterization of seven homologues spanning the sequence space (Figure S5.7). We experimentally validated the predicted tetramer formation by size exclusion chromatography (Table 5.1). Despite the diversity of their active site motifs (VQIY, IQNF, LHIY, and QMIA) in the four homologues predicted to form a DfrB-like tetramer, these homologues exhibit a virtually indistinguishable catalytic efficiency for dihydrofolate reduction. Notably, the three characterized homologues for which the DfrB1-like tetramer was not predicted did not display the ability to reduce dihydrofolate. Interestingly, size exclusion chromatography indicates that these inactive homologues form homomers, suggesting that the DfrB domain has a propensity for multimerization that may go beyond the formation of a DfrB1-like complex.

Table 5.1 Kinetic and oligomeric characterization of DfrB homologues

Cluster	Protein ID	B4 strand motif	K_M^{DHF} (μM)	K_M^{NADPH} (μM)	k_{cat}^{DHF} (s^{-1})	k_{cat}^{DHF}/K_M^{DHF} ($\text{s}^{-1}\mu\text{M}^{-1}$)	TMP resistant	Predicted to form a DfrB1-like tetramer	Multimeric state determined by SEC
<i>I</i>	DfrB1	VQIY	8.2 ± 0.11^a	1.6 ± 0.02^a	0.83 ± 0.01^a	0.10^a	Yes	Yes	Tetramer
<i>I</i>	A0A114V6W3	VQIY	21 ± 7^b	12 ± 1^b	1.1 ± 0.2^b	0.05^b	Yes ^b	Yes	Tetramer and higher order states ^b
Singleton	A0A5E7Z8W7	IQNF	7 ± 4	18 ± 2	1.3 ± 0.03	0.17	Yes	Yes	Tetramer
<i>II</i>	BBD027md-00173 03685	LHIY	28 ± 3	4.8 ± 0.4	2.3 ± 0.1	0.08	Yes	Yes	Tetramer and octomer
<i>II</i>	BBD029md-00326 13692	QMIA	36 ± 4	20 ± 10	4.0 ± 0.3	0.11	Yes	Yes	Tetramer
<i>III</i>	A0A1M7UVF4	LQFL	NS				No	No	Dimer
<i>IV</i>	A0A1R3X8F5	VGYF	NS				No	No	Tetramer
<i>V</i>	A0A345GTK3	TLTL	NS				No	No	Dimer

^a From ³³

^b From ³²

NS: No significant activity detected using from 16 to 84-fold the protein concentration (molar) used to characterize active homologues

A distinguishing feature of tested clinically-relevant members of the DfrB family is their shared thermostability, fully tolerating incubation at 95°C.³² We evaluated this feature by comparing the dihydrofolate reductase activity in *E. coli* lysates following induced expression, before and after heat treatment at 50°C and 75°C. Notably, this *in vitro* lysate assay was less sensitive than the *in vivo* trimethoprim resistance assay, with only 32 of the 114 trimethoprim-resistant homologues showing detectable dihydrofolate reductase activity. Despite having described thermostability as a common feature within the DfrB-protein family following analysis of the clinically-relevant DfrB1 to DfrB7,³² this property

is not conserved in DfrB homologues (Figure 5.4A, Figure S5.8). Thermotolerance is observed in only about half of the DfrB homologues following incubation at 50°C (70% from cluster *I* and 33% from cluster *II*) and in one third following incubation at 75°C (35% from cluster *I* and 33% from cluster *II*), demonstrating that it is not a conserved property in the evolution of the DfrB domain.

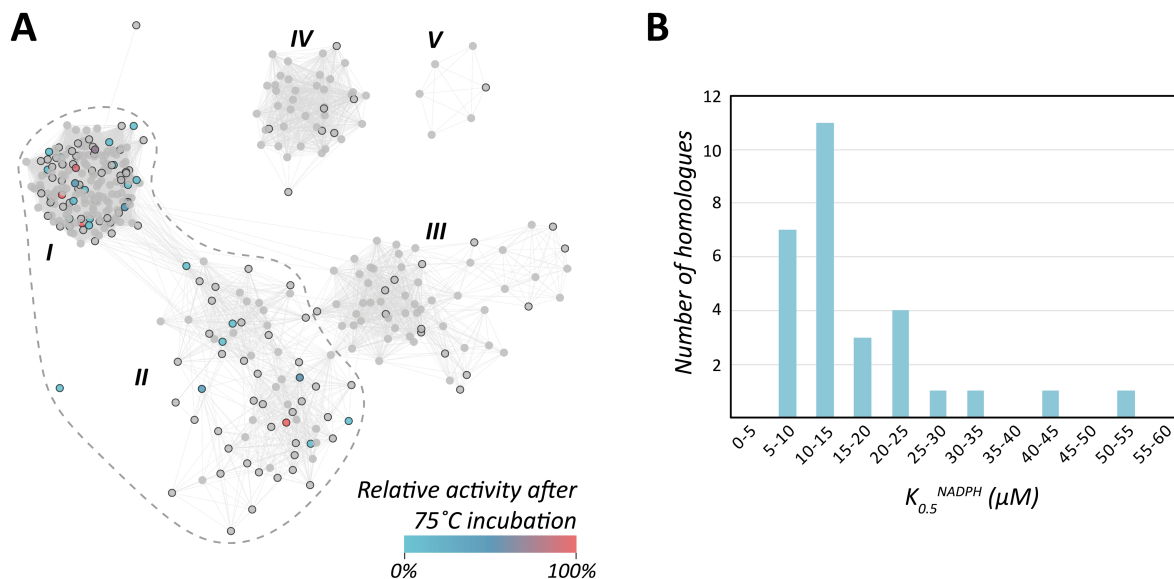


Figure 5.4. Thermostability and dinucleotide binding in the evolution of the DfrB domain.

A. Homologues with detectable dihydrofolate reductase activity in lysate are color-coded according to their thermostability. Following induction of DfrB homologue expression, *E. coli* lysates were subjected to a 10 min incubation at 75°C to determine thermotolerance. All homologues experimentally tested for dihydrofolate reductase activity are indicated with a black border. **B.** The distribution of half-maximal concentration constant for NADPH binding ($K_{0.5}$) for homologues having a detected dihydrofolate reductase activity in lysate. The $K_{0.5}$ value for DfrB1 is 10.8 μM .

Finally, binding to the NADPH cofactor is essential to the dihydrofolate reductase activity of the DfrB enzymes. We characterized the productive affinity of DfrB homologues for this dinucleotide in lysate: this affinity is mostly similar to that of DfrB1, with 83% of the homologues having a $K_{0.5}$ less than 20 μM (Figure 5.4B). Since the *in vivo* concentration of NADPH in exponentially growing *E. coli* is in the range of 120 μM , these homologues would readily bind NADPH in similar conditions.³⁴

Altogether, these data suggest that the DfrB tunnel evolved for ligand binding rather than as a specialized catalyst (Figure 5.3C). The pattern of predicted DfrB1-likier dimer and tetramer formation seen in the SSN (Figure 5.2B,C) is consistent with a monomeric SH3 domain evolving a dimerization interface and a positively-charged ligand-binding surface, in either order. The ligand-binding dimer would then have evolved a second dimerization interface, forming the highly symmetrical homotetramer where the two

equivalent positively-charged binding surfaces line a central tunnel (Figure 5.3C). The formation of a DfrB1-like tetramer, which central tunnel can accommodate one molecule of NADPH and one molecule of DHF, would have led to the catalytic complex.

5.3.4 Investigating the native function of the DfrB domain

Dihydrofolate reduction is essential in most known organisms, as the tetrahydrofolate produced is indispensable for the synthesis of purines and methionine.^{35,36} This reaction is ubiquitously catalyzed by FoaA enzymes, which are monomeric dihydrofolate reductases (Figure S5.9). The FoaA enzymes are 100-fold more efficient at dihydrofolate reduction than DfrB enzymes.³⁷ FoaA is highly specific for the reduction of dihydrofolate. For example, the active site of *E. coli* FoaA forms specific interactions with NADPH and DHF, and conserved residues within this active site increase the pK_a of DHF-N5 from 2.6 to 6.7. This facilitates its solvent-mediated protonation, favoring hydride transfer from NADPH.^{38,39} In comparison, the calculated pK_a of the DHF-N5 for DfrB1 catalysis is 4.5, explaining its poor catalytic efficiency.⁴⁰

Since FoaA are ubiquitous, two orders of magnitude more catalytically efficient at dihydrofolate reduction than DfrB enzymes and have no known natural inhibitors, what evolutionary purpose could have been served by the DfrB homologues prior to the introduction of trimethoprim? What evolutionary purpose could still be served by the homologues identified in environmental bacteria unrelated to human activities, where no selective pressure from trimethoprim exists? We hypothesize that they evolved a different primary function that promiscuously allows dihydrofolate reductase activity.

To gain insight into a putative native function of DfrB homologues, we examined their genomic contexts. Genes with a related function are often organized in operons. As a result, determining the function of proximal genes plays a crucial role in elucidating the function of uncharacterized proteins.⁴¹⁻⁴³ We observed that the homologues predicted to form a DfrB1-like tetramer show a significantly stronger association with genes involved in replication, recombination and DNA repair than homologues where the DfrB1-like tetrameric assembly was not predicted (Figure 5.5A, Figure S5.10A). Among the 139 genomic contexts analyzed, the *dnaN* gene, known for its ring-like clamping action on DNA,⁴⁴ was detected in 17 cases, with a median distance of 2.1 kb from the *dfrB* homologue. Interestingly, *dnaN* was exclusively detected in proximity of *dfrB* predicted to form a DfrB1-like tetramer, such that 27% of those genomic contexts harbor *dnaN* within 10 kb of the *dfrB* homologues. This is consistent with a functional relation. Similarly, the *xerC* gene, a site-specific tyrosine recombinase that is implicated in the resolution of dimeric chromosomes,^{45,46} was found in 23 occurrences at a median distance of 2.0 kb from the DfrB homologous genes. Of these occurrences, 20 were in the context of homologues predicted to form a tetramer and the remaining three were in the context of homologues predicted to form a monomer. While some genomic contexts are conserved between homologues (Figure S5.10B), no clear pattern is observed across all contexts.

Notably, there is a positive correlation between the predicted formation of a DfrB1-like tetramer and the length of the homologues (Figure S5.11). Homotetramerization may have conferred an evolutionary advantage, favoring later evolutionary steps such as gene fusion with functional domains that benefit from the multimerization of the DfrB domain; in this scenario, the native function of the DfrB domain would be to promote protein-protein interactions, as befits an SH3-fold domain. Of note, while the homomultimeric complex is a conserved feature among homologues, no homologue contains more than one DfrB domain in the same peptide chain, nor have *dfrB* gene duplications been observed.

We used Foldseek to identify the function of the other domains fused to the DfrB domains.⁴⁷ Of the 45 predicted proteins in our dataset that are over 160 residues in length, and thus that are likely to contain one or more domains in addition to the 60-residue SH3 fold, seven contain functionally described domains, while many others are alpha-helix bundles of unknown function. Importantly, all seven fusions with a functionally described domain were experimentally confirmed to be functional as dihydrofolate reductases by their trimethoprim resistance phenotype. The fusion of a DfrB domain with a NTP pyrophosphohydrolase predicted to hydrolyze nucleotides is the most common fusion, with three occurrences identified in *Hyphomicrobiales* organisms (Figure 5.5B). These enzymes are reported to dimerize in a canonical manner (Figure S5.12A). A putative NTP pyrophosphohydrolase structure has been experimentally resolved and described to dimerize in a swapped fashion, where the alpha helices are intertwined (Figure S5.12B),⁴⁸ and is predicted in all three NTP pyrophosphohydrolase-DfrB fusions (Figure S5.12C). The negatively charged residues on the side of the NTP pyrophosphohydrolase domain described to bind a magnesium atom are conserved in the three NTP pyrophosphohydrolase-DfrB fusion enzymes. Notably, although these three enzymes share only between 42 and 58 % sequence identity and have widely differing fusion topologies with the DfrB domain, all three genes are identified downstream of a *dnaN* gene, suggesting a functional association (Figure S5.12D).

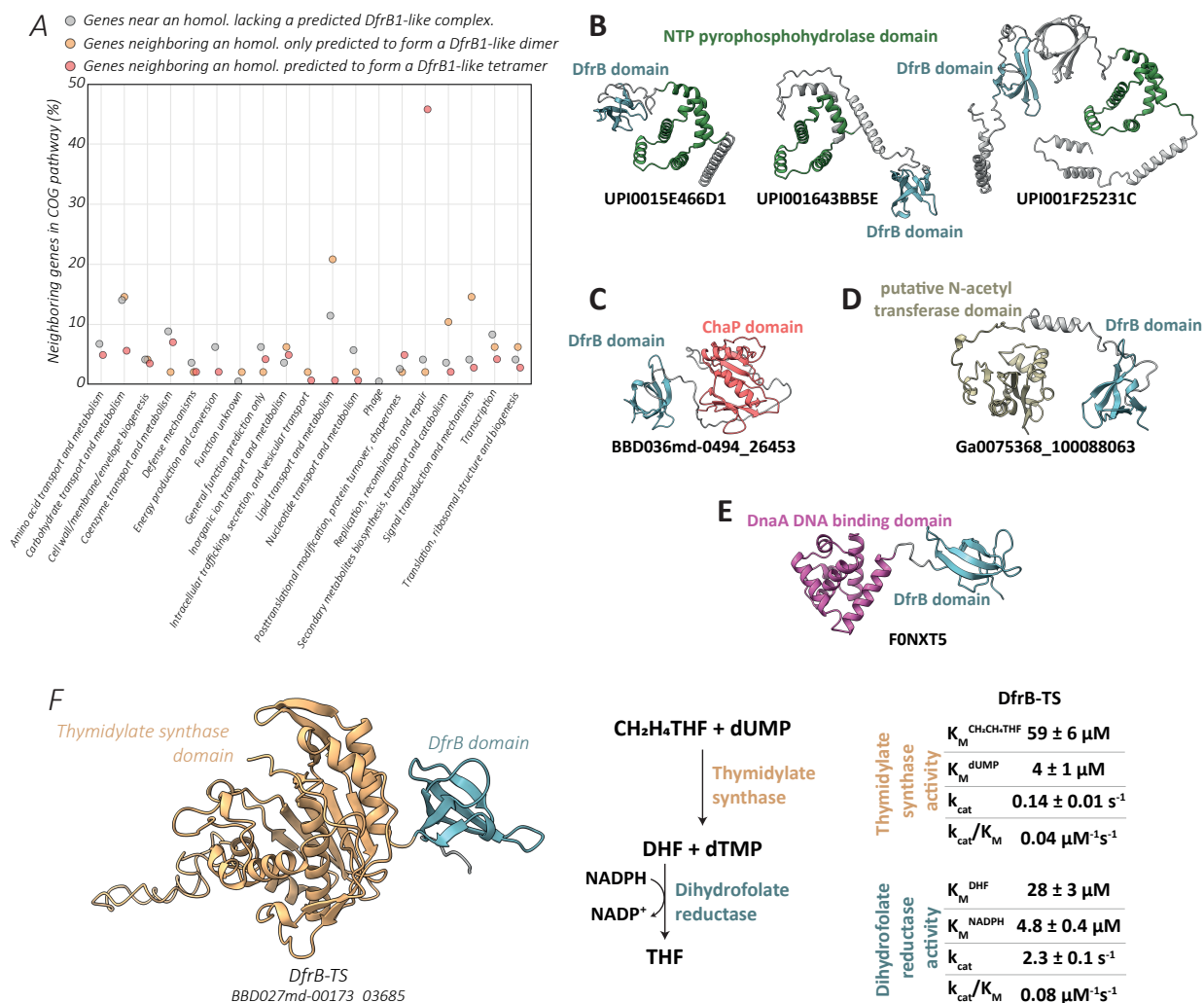


Figure 5.5. Insights into a putative native function of the DfrB domain.

A. Genes within 5 kb of a *dfrB* homologue are annotated according to their corresponding COG pathway and binned according to the AlphaFold-multimer assembly prediction of the proximal DfrB homologue. The DfrB domains fused to **B.** NTP pyrophosphohydrolase domains, **C.** a ChaP oxygenase domain, **D.** a putative N-acetyl transferase domain, **E.** the DNA binding domain of DnaA, or **F.** a thymidylate synthase (TS) domain are represented. The kinetic parameters for the thymidylate synthase and dihydrofolate reductase activities of DfrB-TS in **(F)** are presented.

The DfrB domain was also identified in fusion with a putative ChaP oxygenase domain that catalyzes the final α -pyrone ring formation for the synthesis of chartreusin, a potent antitumor polyketide (Figure 5.5C).⁴⁹ This homodimeric enzyme is active in the presence of FAD, NADH and a flavin reductase.⁴⁹ Residues described to complex with iron are conserved in the ChaP-DfrB protein. Notably, the dimer prediction for this fusion protein simultaneously reproduces the experimentally described dimers of ChaP and DfrB1 (Figure S5.13A). The DfrB domain was also identified in fusion with a putative acetyltransferase domain which are typically described to dimerize: the dimer prediction of this fusion protein is consistent with both the N-acetyltransferase domain and the DfrB domain adopting the reported dimeric conformation (Figure

5.5D, Figure S5.13B).⁵⁰ These domains transfer an acetyl group from acetyl-CoA, which has an adenine nucleotide core. The DfrB domain was also found fused to the DNA-binding domain of DnaA from *Weeksella virosa*. This domain binds to the origin of replication *oriC*, and includes conserved residues responsible for binding to specific DNA sequences (Figure 5.5E, Figure S5.13).⁵¹ Notably, DnaA has been reported to bind cooperatively to DNA, forming large complexes with up to 30 monomers.⁵²

Finally, the DfrB domain was found in fusion with a bacterial thymidylate synthase (TS). We named this fusion protein DfrB-TS. This is notable because TS and dihydrofolate reductase play sequential roles in the folate pathway. This finding strongly supports the hypothesis that the function of the DfrB domain in this fusion is to reduce the dihydrofolate produced by the TS domain (Figure 5.5F). Fusions of TS and dihydrofolate reductase are known, but to date have exclusively involved the fusion of TS with a Fola domain – the ubiquitous, monomeric dihydrofolate reductase. This was first reported in *Crithidia fasciculata* in 1980⁵³ and has since been described in several protozoans and some plants,^{54,55} but never in bacteria as found here (Figure S5.13D). The DfrB-TS and Fola-TS (known as DHFR-TS) fusions thus constitute an elegant example of convergent metabolic evolution, where evolutionarily unrelated domains performing the same catalytic function (DfrB and Fola) are fused to the same TS domain within a metabolic pathway.

Notably, the TS domain functions as a symmetrical dimer.⁵⁶ We experimentally verified that both domains within the DfrB-TS are functional, confirming its bifunctional nature (Figure 5.5F, Figure S5.14). The kinetic parameters of the TS domain are slightly lower than previously characterized TS domains, making its catalytic efficiency 6-fold lower than the TS domain in the DHFR-TS from *Toxoplasma gondii*.⁵⁷

5.4 Discussion

One of the powers of exploring the sequence space of a protein family lies in our ability to uncover the properties that have shaped its evolution. It allows deconvoluting the contribution of functional and biophysical properties to the evolution of protein sequences.^{58–60} This approach is particularly powerful for understudied protein families where sequence space has been characterized asymmetrically due to historical reasons. The DfrB enzymes fall into this category, as only the handful of clinically-relevant DfrB sequences had been characterized prior to the current study. These DfrB are closely related, such that the sequence encoding their SH3 fold is conserved and provides little insight into the evolution of this protein family.

The tetrameric configuration that forms a central and symmetric pore in DfrB1 is the basis of what makes this fold catalytic. It is this precise self-assembly that generates the active site tunnel, which is essential for dihydrofolate reduction (Figure 5.1A).^{22,24} DfrB1 is the only member of the DfrB family for which the homotetrameric structure has been experimentally resolved. Not only is this feature predicted to be

ubiquitous in 274 homologues from clusters *I* and *II* of the sampled sequence space, but it was also revealed to be a sufficient condition for catalysis: of the 123 experimentally characterized homologues predicted to adopt a DfrB1-like tetrameric structure, an impressive 114 (93%) exhibited the ability to reduce dihydrofolate (Figure 5.2A).

The confidence of the AlphaFold-multimer predictions for the homotetramer of these functional homologues is high, as the median self-assessing score is of 0.93 (Figure S5.4). Size exclusion chromatography validated tetramer formation in representative catalytic homologues, confirming successful prediction by AlphaFold-multimer (Table 5.1). Importantly, all 25 homologues for which AlphaFold-multimer did not predict a DfrB1-like tetramer showed no detectable function as dihydrofolate reductases.

To our knowledge, this is the first large-scale experimental evaluation using new experimental data to validate the predictive accuracy of AlphaFold-multimer. The overall performance of this predictive tool for the DfrB enzyme system is remarkable given that the sequences forming both monomer-monomer and dimer-dimer interfaces of catalytically active homologues are not conserved in either residue identity or length (Figure S5.15). For example, although A0A5E7Z8W7 has a seven-residue insertion on a loop at the dimer-dimer interface, this homologue is predicted to form a DfrB1-like tetramer, exhibits the same overall kinetic parameters as all active DfrB domains and assembles into a tetramer in solution (Table 5.1, Figure S5.7B,C).

The relationship between the assembly of the DfrB complex and its catalytic potential is particularly striking; no single residue is shared by all functional homologues, making this a remarkably permissive catalyst (Figure S5.15B). This is most prominent in the B4 strand, which forms the active site tunnel in catalytically functional homologues: 18 demonstrated that different naturally-occurring active-site motifs are catalytically competent. These active sites do not have a distinctive chemical signature: motifs such as VQIY, QMIA, LHVF and CQLW are all capable of reducing DHF with the cofactor NADPH. Therefore, multiple identities of the B4 strand are compatible with the emergence of catalysis in the DfrB domain. However, the electrostatic potential was found to be conserved within homologues capable of forming a DfrB1-like tetramer (Figure 5.3B, Figure S5.6). The tunnels have a positive electrostatic potential: this is consistent with the requirement of these DfrB homologues to bind the negatively charged DHF and NADPH.

The lysine at position 32 in DfrB1 stands out as the most conserved residue among the catalytically active homologues. In the active DfrB homologues, the only substitution observed is arginine, underscoring the

importance of this positive charge (Figure S5.16). The conservation of this residue, known to interact with the NADPH and DHF,⁶¹ highlights the critical role of the positive charge in the catalytic function.

Remarkably, the signature of sequence conservation differs clearly between the clusters (Figure S5.15A). Specifically, both the monomer-monomer and the dimer-dimer interfaces are most highly conserved in cluster *I*. In cluster *II*, the sequence conservation at both interfaces drops considerably, despite DfrB1-like tetramer formation being predicted and demonstrated experimentally. The DfrB1-like monomer-monomer interface is predicted for 54% of the cluster *III* homologues. The B2 strand forming this interface is moderately conserved in these homologues. The B4 strand in these homologues is more variable, the two most common motifs being LQFF and IQIA. The latter motif results in an electropositive surface, which corresponds to half of the surface in the DfrB tunnel (Figure 5.3B). This is consistent with our hypothesis of a DfrB-like dimer having evolved a surface propitious to binding negatively charged molecules (Figure 5.3C).

As mentioned above, the ubiquitous nature of FoaA dihydrofolate reductases suggest that reduction of dihydrofolate may not have been the primary driving force that guided the evolution of the DfrB tetramer. Our results suggest two putative alternative functions for this domain: an oligomerization module and/or a nucleotide binding module. The conservation of the DfrB1-like tunnel throughout clusters *I* and *II* – as opposed to it being an evolutionary event limited to DfrB1 and its closest homologues within cluster *I* – suggests that the tunnel has proven useful and that this property has been recurrently selected for during evolution. Notably, the tetrameric homologues are, on average, significantly longer than those not predicted to form a DfrB1-like tetramer (Figure S5.11); the DfrB domain may have been fused to other domains that benefit from its capacity to oligomerize, or vice versa. This hypothesis gains support from the observation that all domains fused with a DfrB domain for which a function is predicted are recognized to function either as symmetrical dimers or to assemble into higher order complexes (Figure 5.5, Figure S5.12, Figure S5.13). It is unlikely that fusion to additional oligomerizing domains led to DfrB homotetramerization because the fusion of the diverse domains would result from independent events, and most DfrB1-like domains capable of forming a DfrB1-like tetramer are the only domain in the polypeptide chain. Instead, it is plausible that a first homotetramerization event occurred prior to fusions; this is supported by observing that the predicted homodimers do not bear additional functional domains. Therefore, it appears that tetramerization of the DfrB1-like domain is beneficial for the domains it is fused to, potentially to favor their assembly.

A further putative evolutionary advantage stems from the hypothesis that DfrB homologues may broadly bind nucleotide-based molecules. For example, the DfrB domain binding to a nucleotide-based cofactor could make it available for a fused enzyme domain that requires that cofactor. That DfrB homologues

should offer weak binding for nucleotide-based molecules is based firstly on the fact that NADPH and DHF bind simultaneously as substrates, yet the active-site cavity has not evolved great selectivity for these ligands. The tunnel is overly large, including many water molecules bound along with the substrates, which is not characteristic of highly selective binding^{24,62–65}; both NADPH and NADH dinucleotides bind to DfrB1 (K_D of 2.5 and 34 μM , respectively⁶⁶), supporting moderate selectivity. In addition, two NADPH molecules can simultaneously bind within the symmetrical active-site tunnel, with negative cooperativity ($K_{D1} = 2.5 \mu\text{M}$; $K_{D2} = 96 \mu\text{M}$) such that the loosely-bound second NADPH could be available.^{22,66} Here, we report that the apparent affinity of DfrB 30 homologues for NADPH is similar to that of DfrB1 and is consistent with the biological concentrations of NADPH (Figure 5.4B).

Also, the conserved electropositive potential of the tunnels of the homologues predicted to form a DfrB1-like tetramer indicates that this tunnel evolved to bind negatively charged molecules and the conserved lysines at the mouths of the tunnel bind the phosphate groups of NADPH (Figure S5.16).²⁴ This configuration could have evolved to bind additional nucleotide-based molecules. Based on genomic neighborhood analysis, homologues predicted to encode a DfrB1-like tetramer are associated more frequently with genes linked to replication, recombination and repair than their predicted monomeric counterparts. Finally, the domains that we have reported to be fused to the DfrB domain are described to bind nucleotide-based molecules such as nucleotides, DNA, acetyl-CoA, and folates; the DfrB domain homologues may act as reservoirs for those ligands, binding them with moderate specificity to allow their release for use by their fused partner domain. Therefore, DNA, nucleotides, and nucleotide-derived coenzymes represent possible binding partners to the DfrB domain during its evolution.

Three decades of research into characterization of the DfrB1 enzyme have revealed that the reduction of dihydrofolate by the DfrB enzyme results from a convergence of factors.²² First, the relative simplicity of the chemical reaction, involving the reduction of an imine, can be catalyzed by positioning reactive groups in close proximity, without the need for specific chemistry on the part of the enzyme.²³ This is aided by the preprotonation of the imine by the solvent.^{67,68} Secondly, the transition state is achieved by structural exploration facilitated by the disorder of the substrate, rather than by specific interactions stabilized by the enzyme.^{22,37,69} Finally, the catalytic efficiency provided by DfrB enzymes, although low, is biologically sufficient to provide a metabolic advantage, making it an effective catalyst. We demonstrated here that homologues that do not share sequence elements previously thought to be important for the homotetrameric catalyst display similar kinetic parameters to DfrB1, suggesting they might use the same mechanistic strategies for catalysis.

How did the catalytic capacity of the DfrB complex emerge? One might hypothesize that the central pore of the DfrB tetramer evolved its capacity to bind DHF and NADPH after evolving to form the characteristic

DfrB tetramer. This hypothesis would be supported by the existence of homologues that form the DfrB1-like homotetramer but are unable to catalyze the reaction. These proteins would represent an evolutionary intermediate between those unable to form the characteristic tetramer and those that form a DfrB1-like tetramer and catalyze the reduction of dihydrofolate. However, our results are not consistent with such a stepwise evolution, since nearly all (96%; 45 out of 47) of the characterized homologues within cluster *II* predicted to form a DfrB1-like tetramer are functional dihydrofolate reductases.

Rather, our results suggest that the evolution of the dimer-dimer interface would have been the final evolutionary step prior to formation of the catalytically active tetramer. We propose that from a dimer with an electropositive potential at the surface, a second dimerization interface would have evolved to form the tunnel that can accommodate two negatively charged molecules, such as NADPH and DHF (Figure 5.3C). Indeed, we observe a diversity of electrostatic potentials on the B4 strand of the homologues for which DfrB1-like dimer, but not DfrB1-like tetramer, is predicted, some among which exhibit an electropositive potential at the surface that supports the viability of the model (Figure 5.3B). We cannot yet postulate on the evolutionary steps that would have led to this dimer with a surface having electropositive potential. By size exclusion chromatography, we have observed that even the non-catalytically active iterations of the DfrB domain have a propensity to form homomers (Table 5.1). The DfrB1-like dimers observed in cluster *III* would have been the configuration allowing evolution to the DfrB1-like tetramer. This ultimate configuration favored the sequence diversification observed in the homotetrameric and catalytically active populations of clusters *I* and *II*.

Cases of enzymatic activity emerging from a binding protein have rarely been described, but suggest that the chemistry in the binding pocket evolves to accommodate the transition state required for catalysis.^{9,10} The emergence of dihydrofolate reduction catalysis in the DfrB domain is different: reactivity emerges from the formation of an electrostatically positive tunnel environment, where no single chemical signature leads exclusively to catalysis. Our work demonstrates that oligomerization can play a leading role in the formation of an inter-domain surface that can bind and orient molecules in a catalytically active conformation.

5.5 Materials and Methods

5.5.1 Identification of homologues

We used the sensitive homology detection software of HMMER and HH-suite.^{70,71} An MSA consisting of the region corresponding to the SH3 fold in the trimethoprim-resistant DfrB1 to DfrB21 (positions 24 to 78) was used to search UniRef30 (2022_02) with HH-blits, using five iterations. The 367 identified sequences were filtered using HH-filter to retain hits with 85% coverage and 20% sequence identity (-cov 85 -qid 20). The 182 filtered sequences were used as input for hmmsearch, using an E-value of 10^{-3} to

search against UniRef90 (2022_05). In total, 195 sequences were identified. The protein IDs of the hits identified by HH-blits and hmmsearch were used to retrieve the sequences from UniRef. Entries recently removed from the database were not retrieved, resulting in a list of a total of 212 non-redundant homologues.

To explore a larger sequence space, we queried metagenomic data. A total of 2702 metagenomes from the JGI/IMG database (<https://img.jgi.doe.gov/>) were retrieved in May 2020 using a Pfam search for “DHFR_2”.⁷² Filtering of their putative genes using the Pfam keyword “pfam06442” yielded 1524 complete genes starting with a methionine, of which 688 encoded non-redundant proteins. We further increased our exploration by searching microbial single-cell genome database bit-GEM⁷³ with the 182 hh-filtered sequences using hmmsearch with an E-value of 10^{-3} .

We combined the homologues obtained from UniRef, JGI and bit-GEM and filtered the dataset with CD-HIT (85% sequence similarity and 85% length difference cutoff),⁷⁴ for a total of 386 sequences. A total of 148 representative sequences, covering the entire dataset and the diversity of active site motifs identified, were synthesized as N-terminally His-tagged in pET29b by Twist Biosciences.

5.5.2 Generating the sequence similarity network

Briefly, pairwise MMseqs⁷⁵ E-values were calculated between all possible pairs of the 386 representative complete sequences. Pairwise similarities were used to generate a network with Cytoscape⁷⁶ in which a node represents a protein sequence, and an edge represents a pairwise MMseqs E-value. Edges with E-value scores of 10^{-6} and lower are shown.

5.5.3 Generating a DfrB-representative phylogenetic tree

The phylogenetic tree was generated from the MAFFT⁷⁷ alignment of the region of each sequence corresponding to the SH3 fold (positions 24 to 78 in DfrB1) using IQ-TREE⁷⁸ (ultrafast bootstrap analysis, 1000 alignments). The tree was represented using iTOL⁷⁹.

5.5.4 Genomic annotation and analysis

The source of isolation from which each protein was detected was retrieved. The genomic sequences of representative proteins were retrieved using their NCBI sequence ID along with their taxonomic assignment. Metagenomic contigs from JGI and bitBiome were too short for annotation purposes. GC% ratios were calculated for genes whose genomic sequence was at least ten times longer.

From the genomic sequences, 10 kb upstream and downstream of the DfrB homologue gene, for a total of 20 kb, were annotated with Prokka⁸⁰ using the Galaxy server⁸¹. Genomic sequences with shorter genomic

contexts surrounding the DfrB homologue genes were not used. The predicted genes were used as queries against the COG database, and a COG ID was annotated if the blastp E-value was less than 10^{-5} .^{82,83}

COG functional categories were assigned using the COG reference list (<https://www.ncbi.nlm.nih.gov/research/cog/cogcategory/J/>).⁸⁴ In addition to annotating the genomic sequences with PROKKA, we sought to confirm the presence of mobile genetic elements (MGEs) and antibiotic resistance genes (ARGs), as some *dfrB* genes have been previously associated with such elements. For this purpose, we used IntegronFinder⁸⁵ using the Galaxy server⁸¹ and the Resistance Gene Identifier from the Comprehensive Antibiotic Resistance Database (CARD)⁸⁶, respectively. Integrons were identified in contigs using the local detection (--local-max) and search for promoter and attI sites (--promoter-attI) options. Genomic annotations were compiled using R (version 4.1.3) and visualized using the ggplot2⁸⁷ package.

5.5.5 Protein structure prediction

The DfrB1-like dimer and DfrB1-like tetramer for each protein were generated with AlphaFold-multimer⁸⁸ installed on the servers of the University of Haifa, using the region of each sequence corresponding to the SH3 fold. The Mgnify⁸⁹, BFD⁹⁰, UniRef90⁹¹, Uniclust30⁹², UniProt⁹³ and PDB databases were used for these predictions. Out of 25 relaxed models, the model with the highest confidence score (0.8 ipTM [predicted interface TM score] + 0.2 pTM [predicted TM score]) was used for analysis.^{90,94} The ability of the homologues to form a DfrB1-like dimer and a DfrB-like tetramer was determined by generating a contact map from each prediction, as described in Figure S5.3. The full-length sequence of the protein fusions was used to predict homodimeric complexes. The electrostatic potential at the predicted complexes was calculated by the Adaptive Poisson–Boltzmann Solver (APBS) software on PyMOL.⁹⁵

5.5.6 Trimethoprim resistance assay

The *in vivo* trimethoprim resistance assays were performed in triplicates according to⁹⁶ using the agar method. Briefly, *E. coli* BL21(DE3) cells transformed with the homologues were propagated overnight in Luria-Bertani (LB) medium with $50 \mu\text{g mL}^{-1}$ kanamycin. An inoculum of 10^4 colony-forming units (cfu) was spotted on 5% methanol LB agar plates with 0.25 mM IPTG (ThermoFisher) and TMP (Sigma) in two-fold concentration steps up to $600 \mu\text{g mL}^{-1}$; the latter is the highest concentration of TMP soluble in a final concentration of 5% methanol. The TMP concentration inhibiting bacterial growth following 42 h incubation at 37°C was considered to be the minimal inhibitory concentration (MIC). Homologues with a MIC of $75 \mu\text{g/mL}$ and higher are considered to be trimethoprim-resistant.

5.5.7 Thermotolerance

The substrates DHF (synthesized as previously reported⁹⁷) and NADPH (Chem Impex) were quantified spectrophotometrically in 50 mM pH 7 potassium phosphate buffer ($\epsilon_{282\text{nm}}^{\text{DHF}}$ 28 400 M⁻¹cm⁻¹ and $\epsilon_{340\text{nm}}^{\text{NADPH}}$ 6200 M⁻¹cm⁻¹). The lysis of the 148 homologues and the thermotolerance assay of clarified lysates were performed as previously reported in the 4.5.4 section of this thesis.³² Heated lysates were incubated for 10 minutes at 50°C or 75°C.³² Assays were performed in triplicates.

5.5.8 Single curve assay

Inspired by ⁹⁸, complete depletion of NADPH in lysates incubated with DHF was monitored to calculate the half maximum concentration constant ($K_{0.5}$) of DfrB homologues for binding to NADPH as presented in Figure S5.17. First, a calibration curve was generated. Seven reactions using a concentration of NADPH between 3.125 and 50 μM , always with 250 μM DHF and 0.51 μM DfrB1 were followed over 2.5 h. Once every curve was corrected for non-enzymatic substrates degradation, the variation of absorbance over the whole reaction (peak absorbance – lowest absorbance) was calculated for each curve and plotted on a graph in relation to [NADPH]. The corresponding calibration curve, with a R² of 0.98 between the difference in OD and [NADPH], was generated.

Homologues for which significant dihydrofolate reductase activity was detected in lysates were expressed in *E. coli* BL21(DE3) in 96 deep-well plates containing 1 mL of ZYP-5053 medium 20 μL inoculation, and were incubated at 37°C for 3 h, and for 16 h at 22°C, always under agitation. The plates were centrifuged for 30 min, at 4°C, at 2000 g. The cell pellets were resuspended in 300 μL of lysis buffer (100 mM pH 8 potassium phosphate buffer, 10 mM MgSO₄, 1 mM DTT, 0.5 mg/mL lysozyme, 1.5 mM benzamidine, 0.25 mM PMSF, and 0.4 U DNase), agitated for 2 h at 22°C, and then centrifuged. A Beckman Coulter Biomek NXp robot was used to transfer the clarified lysate into 96-well plates. CalA⁹⁹ was used as a negative control for activity, while DfrBH-5³² and DfrB1 were used as positive controls for activity. Lysate concentrations were 40 % (v/v) in 50 μL wells containing 250 μM DHF and 30 μM NADPH in 50 mM pH 7 potassium phosphate buffer. Absorbance depletion at 340 nm was followed during 85 min to capture the totality of the reaction. DHF and NADPH degradation was controlled for by monitoring the depletion of absorbance for both substrate by incubating them individually with DfrB1 lysate. Assays were carried in triplicate. The OD measurements were translated in [NADPH] with the calibration curve, and the data was then fitted in the equation below in Excel using the Solver tool to extract the parameters V_{max} , $K_{0.5}$ and h , the Hill coefficient, a measure of cooperativity.

$$[S]_i = [S]_{i-1} - (t_i - t_{i-1}) \frac{V_{\text{max}}[S]_{i-1}^h}{K_{0.5}^h + [S]_{i-1}^h}$$

5.5.9 Protein purification

Expression of His₆-tagged proteins transformed in *E. coli* BL21(DE3) was carried out as follows. Overnight precultures of 5 mL inoculated 500 mL of Terrific Broth medium containing 50 µg mL⁻¹ kanamycine (Sigma). After initial growth at 37°C to up to OD_{600nm} of 0.6, cells were induced by 1 mM IPTG (Thermo) and expression was carried out at 30°C overnight. The cells were harvested, resuspended in IMAC A buffer (600 mM NaCl, 50 mM Tris, 1 mM CaCl₂, 20 mM imidazole and pH 8) and incubated for 15 min at 4°C with 0.5 mg mL⁻¹ lysozyme (Sigma). The cells were lysed by sonication and centrifuged at 16 000g (Sorvall SLA-3000) at 4°C for 20 min. The supernatant was filtered with a 0.2 µm filter, injected onto a HisTrap FF column (Cytiva) and eluted with IMAC B buffer (600 mM NaCl, 50 mM Tris, 1 mM CaCl₂, 500 mM imidazole and pH 8). Buffer exchange and concentration of protein fractions were carried out with Amicon Ultra Centrifugal Filter Units of either 3K or 30 K molecular weight cut-offs (Fisher), in 50 mM pH 8 potassium phosphate buffer. Pure fractions were pooled together and concentrated. The exact mass was confirmed by the Regional Mass Spectrometry Centre at Université de Montréal.

5.5.10 Kinetic characterization

Kinetic assays were performed in 1 cm pathlength quartz cuvette at 27°C in a Cary 100 Bio UV-Visible (Agilent) spectrophotometer.

For the characterization of the dihydrofolate reductase activity, DHF and NADPH were quantified as described in the section ‘Thermotolerance’. During the assays monitored, the initial rate of linear depletion of NADPH and DHF were monitored in triplicates at 340 nm ($\Delta\epsilon_{340nm}$ 12 300 M⁻¹ cm⁻¹ 100) in 50 mM pH 7 potassium phosphate buffer. For the determination of K_M^{DHF} and K_M^{NADPH} , the concentration range of the variable substrate varied while the second substrate was kept at a concentration of 50 µM.

For the characterization of the thymidylate synthase activity of DfrB-TS, measurements were performed in triplicates in a buffer with 100 mM Tris, 50 mM β-mercaptoethanol and 1 mM EDTA at pH 7.3. The substrate (6R)-5,10-CH₂-H₄folate (CH₂H₄THF) was prepared by incubating for 10 min at room temperature a solution of 2 mM of tetrahydrofolate (THF, Sigma) in 0.038 % formaldehyde (Sigma) and dUMP (Sigma) was prepared in the buffer. The assay was monitored at 340 nm ($\Delta\epsilon_{340nm}$ 6 400 M⁻¹ cm⁻¹ 101) where the production of DHF is detected. For the determination of K_M^{DHF} and K_M^{NADPH} , the concentration range of the variable substrate varied while the second substrate was kept at a concentration of either 40 or 50 µM.

Data were fitted to the Michaelis-Menten equation using nonlinear regression analysis GraphPad Prism version 7 for Mac (GraphPad Software, San Diego, CA, USA). Standard deviation is shown.

For the assay following the bifunctional activity of DfrB-TS, the substrates were prepared in a buffer with 100 mM Tris, 50 mM β-mercaptoethanol and 1 mM EDTA at pH 7.3. Absorbance was monitored at 340

nm for 10 min for different combinations of 80 μM $\text{CH}_2\text{H}_4\text{THF}$, 50 μM NADPH and 50 μM dUMP mixed with the enzyme.

5.5.11 Size exclusion chromatography

The oligomerization states of the DfrB homologues were analysed using analytical SEC with an ÄKTA fast protein liquid chromatography system. The 2.4 mL size exclusion column (Superdex 200 Increase 3.2/300, Cytiva) was calibrated with the Cytiva Gel Filtration Calibration Kit and with lysozyme (Fisher). Injections of 10 μL of proteins at 3 mg/mL were applied onto the column equilibrated with 50 mM potassium phosphate, pH 8, at a flow rate of 0.075 mL min^{-1} . Each protein was injected in triplicates.

5.6 Authors' contributions

C.L.-S.-D.: conceptualization, methodology, investigation, formal analysis, visualization, writing—original draft; S.C.-G.: investigation, formal analysis, visualisation, writing—review and editing; M.S.-A.: investigation, formal analysis; K.I.: investigation; J.N.C.: investigation; S.T.: resources; N.B.T.: conceptualization; R.K.: conceptualization, methodology, software; J.N.P.: conceptualization, supervision, funding acquisition, writing—review and editing.

5.7 Funding

This work was supported by the Natural Science and Engineering Research Council of Canada (NSERC) discovery grant RGPIN-N-2018-04686 and the Canada Research Chair in Engineering of Applied Proteins (J.N.P.). R.K. and N.B.-T. were supported by the Israel Science Foundation (Grant/Award Number: 1764/21). C.L.-S.-D was supported by scholarships from NSERC, Hydro-Québec, APRENTICE and MITACS. S.C.-G. was supported by a scholarship from NSERC.

5.8 Conflicts of interest

K.I. and S.T. were employees of bitBiome, Inc. at the time of the study.

5.9 Acknowledgments

This research is dedicated to the memory of Maxime St-Aubin, our beloved colleague. We thank Samy Cecioni and Andreea R. Schmitzer for providing access to their instruments, and Jeffrey W. Keillor for fruitful insights into single curve kinetic analysis.

5.10 References

- (1) Copley, S. D. Evolution of Efficient Pathways for Degradation of Anthropogenic Chemicals. *Nat. Chem. Biol.* **2009**, 5 (8), 559–566. <https://doi.org/10.1038/nchembio.197>.
- (2) Noor, S.; Taylor, M. C.; Russell, R. J.; Jermin, L. S.; Jackson, C. J.; Oakeshott, J. G.; Scott, C. Intramolecular Epistasis and the Evolution of a New Enzymatic Function. *PLoS ONE* **2012**, 7 (6), e39822. <https://doi.org/10.1371/journal.pone.0039822>.

- (3) Kolvenbach, B. A.; Helbling, D. E.; Kohler, H.-P. E.; Corvini, P. F.-X. Emerging Chemicals and the Evolution of Biodegradation Capacities and Pathways in Bacteria. *Curr. Opin. Biotechnol.* **2014**, *27*, 8–14. <https://doi.org/10.1016/j.copbio.2013.08.017>.
- (4) Arnold, F. H. Design by Directed Evolution. *Acc. Chem. Res.* **1998**, *31* (3), 125–131. <https://doi.org/10.1021/ar960017f>.
- (5) Porter, J. L.; Boon, P. L. S.; Murray, T. P.; Huber, T.; Collyer, C. A.; Ollis, D. L. Directed Evolution of New and Improved Enzyme Functions Using an Evolutionary Intermediate and Multidirectional Search. *ACS Chem. Biol.* **2015**, *10* (2), 611–621. <https://doi.org/10.1021/cb500809f>.
- (6) Chica, R. A.; Doucet, N.; Pelletier, J. N. Semi-Rational Approaches to Engineering Enzyme Activity: Combining the Benefits of Directed Evolution and Rational Design. *Curr. Opin. Biotechnol.* **2005**, *16* (4), 378–384. <https://doi.org/10.1016/j.copbio.2005.06.004>.
- (7) Renata, H.; Wang, Z. J.; Arnold, F. H. Expanding the Enzyme Universe: Accessing Non-Natural Reactions by Mechanism-Guided Directed Evolution. *Angew. Chem. Int. Ed.* **2015**, *54* (11), 3351–3367. <https://doi.org/10.1002/anie.201409470>.
- (8) Chen, K.; Arnold, F. H. Engineering New Catalytic Activities in Enzymes. *Nat. Catal.* **2020**, *3* (3), 203–213. <https://doi.org/10.1038/s41929-019-0385-5>.
- (9) Kaltenbach, M.; Burke, J. R.; Dindo, M.; Pabis, A.; Munsberg, F. S.; Rabin, A.; Kamerlin, S. C. L.; Noel, J. P.; Tawfik, D. S. Evolution of Chalcone Isomerase from a Noncatalytic Ancestor. *Nat. Chem. Biol.* **2018**, *14* (6), 548–555. <https://doi.org/10.1038/s41589-018-0042-3>.
- (10) Clifton, B. E.; Kaczmarek, J. A.; Carr, P. D.; Gerth, M. L.; Tokuriki, N.; Jackson, C. J. Evolution of Cyclohexadienyl Dehydratase from an Ancestral Solute-Binding Protein. *Nat. Chem. Biol.* **2018**, *14* (6), 542–547. <https://doi.org/10.1038/s41589-018-0043-2>.
- (11) Ortmyer, M.; Lafite, P.; Menon, B. R. K.; Tralau, T.; Fisher, K.; Denkhaus, L.; Scrutton, N. S.; Rigby, S. E. J.; Munro, A. W.; Hay, S.; Leys, D. An Oxidative N-Demethylase Reveals PAS Transition from Ubiquitous Sensor to Enzyme. *Nature* **2016**, *539* (7630), 593–597. <https://doi.org/10.1038/nature20159>.
- (12) Tam, R.; Saier, M. H. A Bacterial Periplasmic Receptor Homologue with Catalytic Activity: Cyclohexadienyl Dehydratase of *Pseudomonas Aeruginosa* Is Homologous to Receptors Specific for Polar Amino Acids. *Res. Microbiol.* **1993**, *144* (3), 165–169. [https://doi.org/10.1016/0923-2508\(93\)90041-Y](https://doi.org/10.1016/0923-2508(93)90041-Y).
- (13) Kaur, G.; Subramanian, S. Repurposing TRASH: Emergence of the Enzyme Organomercurial Lyase from a Non-Catalytic Zinc Finger Scaffold. *J. Struct. Biol.* **2014**, *188* (1), 16–21. <https://doi.org/10.1016/j.jsb.2014.09.001>.
- (14) Harms, M. J. Enzymes Emerge by Upcycling. *Nat. Chem. Biol.* **2018**, *14* (6), 526–527. <https://doi.org/10.1038/s41589-018-0064-x>.
- (15) Alvarez-Carreño, C.; Penev, P. I.; Petrov, A. S.; Williams, L. D. Fold Evolution before LUCA: Common Ancestry of SH3 Domains and OB Domains. *Mol. Biol. Evol.* **2021**, *38* (11), 5134–5143. <https://doi.org/10.1093/molbev/msab240>.
- (16) Dionne, U.; Percival, L. J.; Chartier, F. J. M.; Landry, C. R.; Bisson, N. SRC Homology 3 Domains: Multifaceted Binding Modules. *Trends Biochem. Sci.* **2022**, *47* (9), 772–784. <https://doi.org/10.1016/j.tibs.2022.04.005>.

- (17) Kishan, K.; Agrawal, V. SH3-like Fold Proteins Are Structurally Conserved and Functionally Divergent. *Curr. Protein Pept. Sci.* **2005**, *6* (2), 143–150. <https://doi.org/10.2174/1389203053545444>.
- (18) Mager, P. P. The Active Site of HIV-1 Protease. *Med. Res. Rev.* **2001**, *21* (4), 348–353. <https://doi.org/10.1002/med.1012>.
- (19) Radha Kishan, K. V.; Scita, G.; Wong, W. T.; Di Fiore, P. P.; Newcomer, M. E. The SH3 Domain of Eps8 Exists as a Novel Intertwined Dimer. *Nat. Struct. Biol.* **1997**, *4* (9), 739–743. <https://doi.org/10.1038/nsb0997-739>.
- (20) Delbrück, H.; Ziegelin, G.; Lanka, E.; Heinemann, U. An Src Homology 3-like Domain Is Responsible for Dimerization of the Repressor Protein KorB Encoded by the Promiscuous IncP Plasmid RP4. *J. Biol. Chem.* **2002**, *277* (6), 4191–4198. <https://doi.org/10.1074/jbc.M110103200>.
- (21) Cheng, H.; Schaeffer, R. D.; Liao, Y.; Kinch, L. N.; Pei, J.; Shi, S.; Kim, B.-H.; Grishin, N. V. ECOD: An Evolutionary Classification of Protein Domains. *PLoS Comput. Biol.* **2014**, *10* (12), e1003926. <https://doi.org/10.1371/journal.pcbi.1003926>.
- (22) Lemay-St-Denis, C.; Pelletier, J. N. From a Binding Module to Essential Catalytic Activity: How Nature Stumbled on a Good Thing. *Chem. Commun.* **2023**, *59* (84), 12560–12572. <https://doi.org/10.1039/D3CC04209J>.
- (23) Schmitzer, A. R.; Lépine, F.; Pelletier, J. N. Combinatorial Exploration of the Catalytic Site of a Drug-Resistant Dihydrofolate Reductase: Creating Alternative Functional Configurations. *Protein Eng. Des. Sel.* **2004**, *17* (11), 809–819. <https://doi.org/10.1093/protein/gzh090>.
- (24) Krahn, J. M.; Jackson, M. R.; DeRose, E. F.; Howell, E. E.; London, R. E. Crystal Structure of a Type II Dihydrofolate Reductase Catalytic Ternary Complex[†]. *Biochemistry* **2007**, *46* (51), 14878–14888. <https://doi.org/10.1021/bi701532r>.
- (25) Duff, M. R.; Chopra, S.; Strader, M. B.; Agarwal, P. K.; Howell, E. E. Tales of Dihydrofolate Binding to R67 Dihydrofolate Reductase. *Biochemistry* **2016**, *55* (1), 133–145. <https://doi.org/10.1021/acs.biochem.5b00981>.
- (26) Alonso, H.; Gready, J. E. Integron-Sequestered Dihydrofolate Reductase: A Recently Redeployed Enzyme. *Trends Microbiol.* **2006**, *14* (5), 236–242. <https://doi.org/10.1016/j.tim.2006.03.003>.
- (27) Lemay-St-Denis, C.; Diwan, S.-S.; Pelletier, J. N. The Bacterial Genomic Context of Highly Trimethoprim-Resistant DfrB Dihydrofolate Reductases Highlights an Emerging Threat to Public Health. *Antibiotics* **2021**, *10* (4), 433. <https://doi.org/10.3390/antibiotics10040433>.
- (28) Kneis, D.; Lemay-St-Denis, C.; Cellier-Goetghebeur, S.; Elena, A. X.; Berendonk, T. U.; Pelletier, J. N.; Heß, S. Trimethoprim Resistance in Surface and Wastewater Is Mediated by Contrasting Variants of the dfrB Gene. *ISME J.* **2023**, *17* (9), 1455–1466. <https://doi.org/10.1038/s41396-023-01460-7>.
- (29) Cellier-Goetghebeur, S.; Lafontaine, K.; Lemay-St-Denis, C.; Tsamo, P.; Bonneau-Burke, A.; Copp, J. N.; Pelletier, J. N. Discovery of Highly Trimethoprim-Resistant DfrB Dihydrofolate Reductases in Diverse Environmental Settings Suggests an Evolutionary Advantage Unrelated to Antibiotic Resistance. *Antibiotics* **2022**, *11* (12), 1768. <https://doi.org/10.3390/antibiotics11121768>.
- (30) Pham, D. N.; Wu, Q.; Li, M. Global Profiling of Antibiotic Resistomes in Maize Rhizospheres. *Arch. Microbiol.* **2023**, *205* (3), 89. <https://doi.org/10.1007/s00203-023-03424-z>.

- (31) Kneis, D.; Berendonk, T. U.; Forslund, S. K.; Hess, S. Antibiotic Resistance Genes in River Biofilms: A Metagenomic Approach toward the Identification of Sources and Candidate Hosts. *Environ. Sci. Technol.* **2022**, *56* (21), 14913–14922. <https://doi.org/10.1021/acs.est.2c00370>.
- (32) Lemay-St-Denis, C.; Alejaldre, L.; Jemouai, Z.; Lafontaine, K.; St-Aubin, M.; Hitache, K.; Valikhani, D.; Weerasinghe, N. W.; Létourneau, M.; Thibodeaux, C. J.; Doucet, N.; Baron, C.; Copp, J. N.; Pelletier, J. N. A Conserved SH3-like Fold in Diverse Putative Proteins Tetramerizes into an Oxidoreductase Providing an Antimicrobial Resistance Phenotype. *Philos. Trans. R. Soc. B Biol. Sci.* **2023**, *378* (1871), 20220040. <https://doi.org/10.1098/rstb.2022.0040>.
- (33) Toulouse, J. L.; Shi, G.; Lemay-St-Denis, C.; Ebert, M. C. C. J. C.; Deon, D.; Gagnon, M.; Ruediger, E.; Saint-Jacques, K.; Forge, D.; Vanden Eynde, J. J.; Marinier, A.; Ji, X.; Pelletier, J. N. Dual-Target Inhibitors of the Folate Pathway Inhibit Intrinsically Trimethoprim-Resistant DfrB Dihydrofolate Reductases. *ACS Med. Chem. Lett.* **2020**, *11* (11), 2261–2267. <https://doi.org/10.1021/acsmchemlett.0c00393>.
- (34) Bennett, B. D.; Kimball, E. H.; Gao, M.; Osterhout, R.; Van Dien, S. J.; Rabinowitz, J. D. Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in Escherichia Coli. *Nat. Chem. Biol.* **2009**, *5* (8), 593–599. <https://doi.org/10.1038/nchembio.186>.
- (35) Fox, J. T.; Stover, P. J. Chapter 1 Folate-Mediated One-Carbon Metabolism. In *Vitamins & Hormones*; Elsevier, 2008; Vol. 79, pp 1–44. [https://doi.org/10.1016/S0083-6729\(08\)00401-9](https://doi.org/10.1016/S0083-6729(08)00401-9).
- (36) Myllykallio, H.; Leduc, D.; Filee, J.; Liebl, U. Life without Dihydrofolate Reductase FoaA. *Trends Microbiol.* **2003**, *11* (5), 220–223. [https://doi.org/10.1016/S0966-842X\(03\)00101-X](https://doi.org/10.1016/S0966-842X(03)00101-X).
- (37) Howell, E. E. Searching Sequence Space: Two Different Approaches to Dihydrofolate Reductase Catalysis. *ChemBioChem* **2005**, *6* (4), 590–600. <https://doi.org/10.1002/cbic.200400237>.
- (38) Liu, C. T.; Francis, K.; Layfield, J. P.; Huang, X.; Hammes-Schiffer, S.; Kohen, A.; Benkovic, S. J. Escherichia Coli Dihydrofolate Reductase Catalyzed Proton and Hydride Transfers: Temporal Order and the Roles of Asp27 and Tyr100. *Proc. Natl. Acad. Sci.* **2014**, *111* (51), 18231–18236. <https://doi.org/10.1073/pnas.1415940111>.
- (39) Bystroff, C.; Oatley, S. J.; Kraut, J. Crystal Structures of Escherichia Coli Dihydrofolate Reductase: The NADP+ Holoenzyme and the Folate .Cntdot. NADP+ Ternary Complex. Substrate Binding and a Model for the Transition State. *Biochemistry* **1990**, *29* (13), 3263–3277. <https://doi.org/10.1021/bi00465a018>.
- (40) Mhashal, A. R.; Pshetitsky, Y.; Cheatum, C. M.; Kohen, A.; Major, D. T. Evolutionary Effects on Bound Substrate p K_a in Dihydrofolate Reductase. *J. Am. Chem. Soc.* **2018**, *140* (48), 16650–16660. <https://doi.org/10.1021/jacs.8b09089>.
- (41) Pereira, J. GCsnap: Interactive Snapshots for the Comparison of Protein-Coding Genomic Contexts. *J. Mol. Biol.* **2021**, *433* (11), 166943. <https://doi.org/10.1016/j.jmb.2021.166943>.
- (42) Dandekar, T. Conservation of Gene Order: A Fingerprint of Proteins That Physically Interact. *Trends Biochem. Sci.* **1998**, *23* (9), 324–328. [https://doi.org/10.1016/S0968-0004\(98\)01274-2](https://doi.org/10.1016/S0968-0004(98)01274-2).
- (43) Makarova, K. S.; Wolf, Y. I.; Koonin, E. V. Towards Functional Characterization of Archaeal Genomic Dark Matter. *Biochem. Soc. Trans.* **2019**, *47* (1), 389–398. <https://doi.org/10.1042/BST20180560>.
- (44) Burnouf, D. Y.; Olieric, V.; Wagner, J.; Fujii, S.; Reinbolt, J.; Fuchs, R. P. P.; Dumas, P. Structural

- and Biochemical Analysis of Sliding Clamp/Ligand Interactions Suggest a Competition Between Replicative and Translesion DNA Polymerases. *J. Mol. Biol.* **2004**, 335 (5), 1187–1197. <https://doi.org/10.1016/j.jmb.2003.11.049>.
- (45) Sherratt, D. J.; Søballe, B.; Barre, F.; Filipe, S.; Lau, I.; Massey, T.; Yates, J. Recombination and Chromosome Segregation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2004**, 359 (1441), 61–69. <https://doi.org/10.1098/rstb.2003.1365>.
- (46) Blakely, G.; May, G.; McCulloch, R.; Arciszewska, L. K.; Burke, M.; Lovett, S. T.; Sherratt, D. J. Two Related Recombinases Are Required for Site-Specific Recombination at Dif and Cer in *E. Coli* K12. *Cell* **1993**, 75 (2), 351–361. [https://doi.org/10.1016/0092-8674\(93\)80076-Q](https://doi.org/10.1016/0092-8674(93)80076-Q).
- (47) Van Kempen, M.; Kim, S. S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C. L. M.; Söding, J.; Steinegger, M. Fast and Accurate Protein Structure Search with Foldseek. *Nat. Biotechnol.* **2023**. <https://doi.org/10.1038/s41587-023-01773-0>.
- (48) Han, G. W.; Elsliger, M.-A.; Yeates, T. O.; Xu, Q.; Murzin, A. G.; Krishna, S. S.; Jaroszewski, L.; Abdubek, P.; Astakhova, T.; Axelrod, H. L.; et al. Structure of a Putative NTP Pyrophosphohydrolase: YP_001813558.1 from *Exiguobacterium Sibiricum* 255-15. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.* **2010**, 66 (10), 1237–1244. <https://doi.org/10.1107/S1744309110025534>.
- (49) Wang, Y. S.; Zhang, B.; Zhu, J.; Yang, C. L.; Guo, Y.; Liu, C. L.; Liu, F.; Huang, H.; Zhao, S.; Liang, Y.; Jiao, R. H.; Tan, R. X.; Ge, H. M. Molecular Basis for the Final Oxidative Rearrangement Steps in Chartreusin Biosynthesis. *J. Am. Chem. Soc.* **2018**, 140 (34), 10909–10914. <https://doi.org/10.1021/jacs.8b06623>.
- (50) Vetting, M. W.; S. De Carvalho, L. P.; Yu, M.; Hegde, S. S.; Magnet, S.; Roderick, S. L.; Blanchard, J. S. Structure and Functions of the GNAT Superfamily of Acetyltransferases. *Arch. Biochem. Biophys.* **2005**, 433 (1), 212–226. <https://doi.org/10.1016/j.abb.2004.09.003>.
- (51) Tsodikov, O. V.; Biswas, T. Structural and Thermodynamic Signatures of DNA Recognition by Mycobacterium Tuberculosis DnaA. *J. Mol. Biol.* **2011**, 410 (3), 461–476. <https://doi.org/10.1016/j.jmb.2011.05.007>.
- (52) Fuller, R. S.; Funnell, B. E.; Kornberg, A. The dnaA Protein Complex with the *E. Coli* Chromosomal Replication Origin (oriC) and Other DNA Sites. *Cell* **1984**, 38 (3), 889–900. [https://doi.org/10.1016/0092-8674\(84\)90284-8](https://doi.org/10.1016/0092-8674(84)90284-8).
- (53) Ferone, R.; Roland, S. Dihydrofolate Reductase: Thymidylate Synthase, a Bifunctional Polypeptide from *Crithidia Fasciculata*. *Proc. Natl. Acad. Sci.* **1980**, 77 (10), 5802–5806. <https://doi.org/10.1073/pnas.77.10.5802>.
- (54) Ivanetich, K. M.; Santi, D. V. Bifunctional Thymidylate Synthase-dihydrofolate Reductase in Protozoa. *FASEB J.* **1990**, 4 (6), 1591–1597. <https://doi.org/10.1096/fasebj.4.6.2180768>.
- (55) Balestrazzi, A.; Branzoni, M.; Carbonera, D.; Parisi, B.; Cella, R. Biochemical Evidence for the Presence of a Bifunctional Dihydrofolate Reductase-Thymidylate Synthase in Plant Species. *J. Plant Physiol.* **1995**, 147 (2), 263–266. [https://doi.org/10.1016/S0176-1617\(11\)81515-4](https://doi.org/10.1016/S0176-1617(11)81515-4).
- (56) Carreras, C. W.; Santi, D. V. The Catalytic Mechanism and Structure of Thymidylate Synthase. *Annu. Rev. Biochem.* **1995**, 64 (1), 721–762. <https://doi.org/10.1146/annurev.bi.64.070195.003445>.
- (57) Trujillo, M.; Donald, R. G. K.; Roos, D. S.; Greene, P. J.; Santi, D. V. Heterologous Expression and

- Characterization of the Bifunctional Dihydrofolate Reductase–Thymidylate Synthase Enzyme of *Toxoplasma Gondii*. *Biochemistry* **1996**, 35 (20), 6366–6374. <https://doi.org/10.1021/bi952923q>.
- (58) Akiva, E.; Copp, J. N.; Tokuriki, N.; Babbitt, P. C. Evolutionary and Molecular Foundations of Multiple Contemporary Functions of the Nitroreductase Superfamily. *Proc. Natl. Acad. Sci.* **2017**, 114 (45), E9549–E9558. <https://doi.org/10.1073/pnas.1706849114>.
- (59) Durairaj, J.; Waterhouse, A. M.; Mets, T.; Brodiazhenko, T.; Abdullah, M.; Studer, G.; Tauriello, G.; Akdel, M.; Andreeva, A.; Bateman, A.; Tenson, T.; Hauryliuk, V.; Schwede, T.; Pereira, J. Uncovering New Families and Folds in the Natural Protein Universe. *Nature* **2023**, 622 (7983), 646–653. <https://doi.org/10.1038/s41586-023-06622-3>.
- (60) Copp, J. N.; Akiva, E.; Babbitt, P. C.; Tokuriki, N. Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry* **2018**, 57 (31), 4651–4662. <https://doi.org/10.1021/acs.biochem.8b00473>.
- (61) Hicks, S. N.; Smiley, R. D.; Stinnett, L. G.; Minor, K. H.; Howell, E. E. Role of Lys-32 Residues in R67 Dihydrofolate Reductase Probed by Asymmetric Mutations. *J. Biol. Chem.* **2004**, 279 (45), 46995–47002. <https://doi.org/10.1074/jbc.M404484200>.
- (62) Chopra, S.; Dooling, R. M.; Horner, C. G.; Howell, E. E. A Balancing Act between Net Uptake of Water during Dihydrofolate Binding and Net Release of Water upon NADPH Binding in R67 Dihydrofolate Reductase. *J. Biol. Chem.* **2008**, 283 (8), 4690–4698. <https://doi.org/10.1074/jbc.M709443200>.
- (63) Chopra, S.; Lynch, R.; Kim, S.-H.; Jackson, M.; Howell, E. E. Effects of Temperature and Viscosity on R67 Dihydrofolate Reductase Catalysis. *Biochemistry* **2006**, 45 (21), 6596–6605. <https://doi.org/10.1021/bi052504l>.
- (64) Alonso, H.; Gillies, M. B.; Cummins, P. L.; Bliznyuk, A. A.; Gready, J. E. Multiple Ligand-Binding Modes in Bacterial R67 Dihydrofolate Reductase. *J. Comput. Aided Mol. Des.* **2005**, 19 (3), 165–187. <https://doi.org/10.1007/s10822-005-3693-6>.
- (65) Mhashal, A. R.; Major, D. T. Temperature-Dependent Kinetic Isotope Effects in R67 Dihydrofolate Reductase from Path-Integral Simulations. *J. Phys. Chem. B* **2021**, 125 (5), 1369–1377. <https://doi.org/10.1021/acs.jpcc.0c10318>.
- (66) Jackson, M.; Chopra, S.; Smiley, R. D.; Maynard, P. O.; Rosowsky, A.; London, R. E.; Levy, L.; Kalman, T. I.; Howell, E. E. Calorimetric Studies of Ligand Binding in R67 Dihydrofolate Reductase. *Biochemistry* **2005**, 44 (37), 12420–12433. <https://doi.org/10.1021/bi050881s>.
- (67) Narayana, N.; Matthews, D. A.; Howell, E. E.; Xuong, N. A Plasmid-Encoded Dihydrofolate Reductase from Trimethoprim-Resistant Bacteria Has a Novel D2-Symmetric Active Site. *Nat. Struct. Mol. Biol.* **1995**, 2 (11), 1018–1025. <https://doi.org/10.1038/nsb1195-1018>.
- (68) Park, H.; Zhuang, P.; Nichols, R.; Howell, E. E. Mechanistic Studies of R67 Dihydrofolate Reductase. *J. Biol. Chem.* **1997**, 272 (4), 2252–2258. <https://doi.org/10.1074/jbc.272.4.2252>.
- (69) Stinnett, L. G.; Smiley, R. D.; Hicks, S. N.; Howell, E. E. “Catch 222,” the Effects of Symmetry on Ligand Binding and Catalysis in R67 Dihydrofolate Reductase as Determined by Mutations at Tyr-69. *J. Biol. Chem.* **2004**, 279 (45), 47003–47009. <https://doi.org/10.1074/jbc.M404485200>.
- (70) Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, 7 (10), e1002195. <https://doi.org/10.1371/journal.pcbi.1002195>.

- (71) Steinegger, M.; Meier, M.; Mirdita, M.; Vöhringer, H.; Haunsberger, S. J.; Söding, J. HH-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation. *BMC Bioinformatics* **2019**, *20* (1), 473. <https://doi.org/10.1186/s12859-019-3019-7>.
- (72) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D412–D419. <https://doi.org/10.1093/nar/gkaa913>.
- (73) Nishikawa, Y.; Kogawa, M.; Hosokawa, M.; Wagatsuma, R.; Mineta, K.; Takahashi, K.; Ide, K.; Yura, K.; Behzad, H.; Gojobori, T.; Takeyama, H. Validation of the Application of Gel Beads-Based Single-Cell Genome Sequencing Platform to Soil and Seawater. *ISME Commun.* **2022**, *2* (1), 92. <https://doi.org/10.1038/s43705-022-00179-4>.
- (74) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28* (23), 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>.
- (75) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35* (11), 1026–1028. <https://doi.org/10.1038/nbt.3988>.
- (76) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504. <https://doi.org/10.1101/gr.1239303>.
- (77) Katoh, K.; Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30* (4), 772–780. <https://doi.org/10.1093/molbev/mst010>.
- (78) Nguyen, L.-T.; Schmidt, H. A.; von Haeseler, A.; Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **2015**, *32* (1), 268–274. <https://doi.org/10.1093/molbev/msu300>.
- (79) Letunic, I.; Bork, P. Interactive Tree of Life (iTOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees. *Nucleic Acids Res.* **2016**, *44* (W1), W242–245. <https://doi.org/10.1093/nar/gkw290>.
- (80) Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* **2014**, *30* (14), 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>.
- (81) The Galaxy Community; Afgan, E.; Nekrutenko, A.; Grünig, B. A.; Blankenberg, D.; Goecks, J.; Schatz, M. C.; Ostrovsky, A. E.; Mahmoud, A.; Lonie, A. J.; et al. The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2022 Update. *Nucleic Acids Res.* **2022**, *50* (W1), W345–W351. <https://doi.org/10.1093/nar/gkac247>.
- (82) Altschul, S. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389–3402. <https://doi.org/10.1093/nar/25.17.3389>.
- (83) Galperin, M. Y.; Wolf, Y. I.; Makarova, K. S.; Vera Alvarez, R.; Landsman, D.; Koonin, E. V. COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens. *Nucleic Acids Res.* **2021**, *49* (D1), D274–D281. <https://doi.org/10.1093/nar/gkaa1018>.
- (84) Tatusov, R. L.; Koonin, E. V.; Lipman, D. J. A Genomic Perspective on Protein Families. *Science* **1997**, *278* (5338), 631–637. <https://doi.org/10.1126/science.278.5338.631>.

- (85) Néron, B.; Littner, E.; Haudiquet, M.; Perrin, A.; Cury, J.; Rocha, E. IntegronFinder 2.0: Identification and Analysis of Integrations across Bacteria, with a Focus on Antibiotic Resistance in *Klebsiella*. *Microorganisms* **2022**, *10* (4), 700. <https://doi.org/10.3390/microorganisms10040700>.
- (86) Alcock, B. P.; Raphenya, A. R.; Lau, T. T. Y.; Tsang, K. K.; Bouchard, M.; Edalatmand, A.; Huynh, W.; Nguyen, A.-L. V.; Cheng, A. A.; Liu, S.; et al. CARD 2020: Antibiotic Resistance Surveillance with the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **2019**, gkz935. <https://doi.org/10.1093/nar/gkz935>.
- (87) Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*, Second edition.; Use R!; Springer: Switzerland, 2016. <https://doi.org/10.1007/978-3-319-24277-4>.
- (88) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Židek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. *Protein Complex Prediction with AlphaFold-Multimer*; preprint; Bioinformatics, 2021. <https://doi.org/10.1101/2021.10.04.463034>.
- (89) Richardson, L.; Allen, B.; Baldi, G.; Beracochea, M.; Bileschi, M. L.; Burdett, T.; Burgin, J.; Caballero-Pérez, J.; Cochrane, G.; Colwell, L. J.; et al. MGnify: The Microbiome Sequence Data Analysis Resource in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D753–D759. <https://doi.org/10.1093/nar/gkac1080>.
- (90) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- (91) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; the UniProt Consortium. UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics* **2015**, *31* (6), 926–932. <https://doi.org/10.1093/bioinformatics/btu739>.
- (92) Mirdita, M.; von den Driesch, L.; Galiez, C.; Martin, M. J.; Söding, J.; Steinegger, M. Uniclust Databases of Clustered and Deeply Annotated Protein Sequences and Alignments. *Nucleic Acids Res.* **2017**, *45* (D1), D170–D176. <https://doi.org/10.1093/nar/gkw1081>.
- (93) The UniProt Consortium; Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; et al. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523–D531. <https://doi.org/10.1093/nar/gkac1052>.
- (94) Wallner, B. *AFsample: Improving Multimer Prediction with AlphaFold Using Aggressive Sampling*; preprint; Bioinformatics, 2022. <https://doi.org/10.1101/2022.12.20.521205>.
- (95) Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L. E.; Brookes, D. H.; Wilson, L.; Chen, J.; Liles, K.; et al. Improvements to the APBS Biomolecular Solvation Software Suite. *Protein Sci.* **2018**, *27* (1), 112–128. <https://doi.org/10.1002/pro.3280>.
- (96) Wiegand, I.; Hilpert, K.; Hancock, R. E. W. Agar and Broth Dilution Methods to Determine the Minimal Inhibitory Concentration (MIC) of Antimicrobial Substances. *Nat. Protoc.* **2008**, *3* (2), 163–175. <https://doi.org/10.1038/nprot.2007.521>.
- (97) Blakley, R. L. Crystalline Dihydropteroylglutamic Acid. *Nature* **1960**, *188* (4746), 231–232. <https://doi.org/10.1038/188231a0>.
- (98) Tamer, Y. T.; Gaszek, I. K.; Abdizadeh, H.; Batur, T. A.; Reynolds, K. A.; Atilgan, A. R.; Atilgan, C.; Toprak, E. High-Order Epistasis in Catalytic Power of Dihydrofolate Reductase Gives Rise to a

Rugged Fitness Landscape in the Presence of Trimethoprim Selection. *Mol Biol Evol* **2019**, *36* (7), 1533–1550. <https://doi.org/10.1093/molbev/msz086>.

- (99) Alejaldre, L.; Lemay-St-Denis, C.; Pelletier, J. N.; Quaglia, D. Tuning Selectivity in CalA Lipase: Beyond Tunnel Engineering. *Biochemistry* **2023**, *62* (2), 396–409. <https://doi.org/10.1021/acs.biochem.2c00513>.
- (100) Baccanari, D.; Phillips, A.; Smith, S.; Sinski, D.; Burchall, J. Purification and Properties of *Escherichia Coli* Dihydrofolate Reductase. *Biochemistry* **1975**, *14* (24), 5267–5273. <https://doi.org/10.1021/bi00695a006>.
- (101) Spencer, H. T.; Villafranca, J. E.; Appleman, J. R. Kinetic Scheme for Thymidylate Synthase from *Escherichia Coli*: Determination from Measurements of Ligand Binding, Primary and Secondary Isotope Effects, and Pre-Steady-State Catalysis. *Biochemistry* **1997**, *36* (14), 4212–4222. <https://doi.org/10.1021/bi961794q>.

5.11 Supporting information

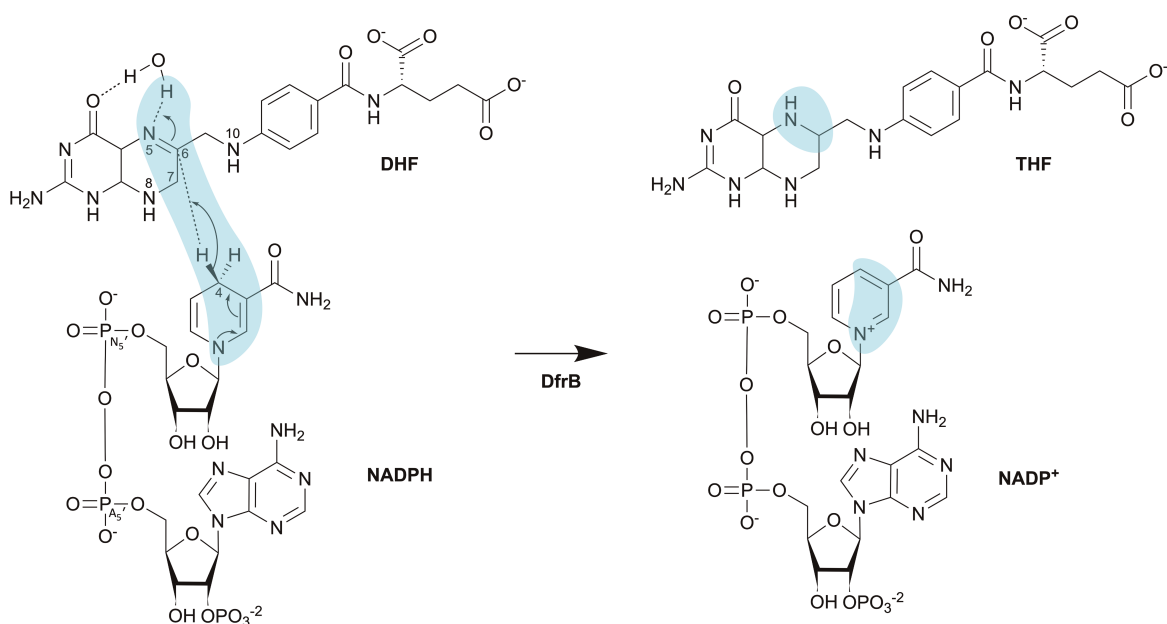


Figure S5.1. The DfrB enzymes catalyze the hydride transfer from NADPH to DHF.

Following protonation of DHF N5 by the solvent, the NADPH hydride is transferred to the DHF C6. The product, the coenzyme tetrahydrofolate, serves as an essential carrier of ‘one-carbon units’ in core metabolic reactions that include purine biosynthesis. Chemical groups implicated in the hydride transfer are highlighted in blue. Adapted from ¹.

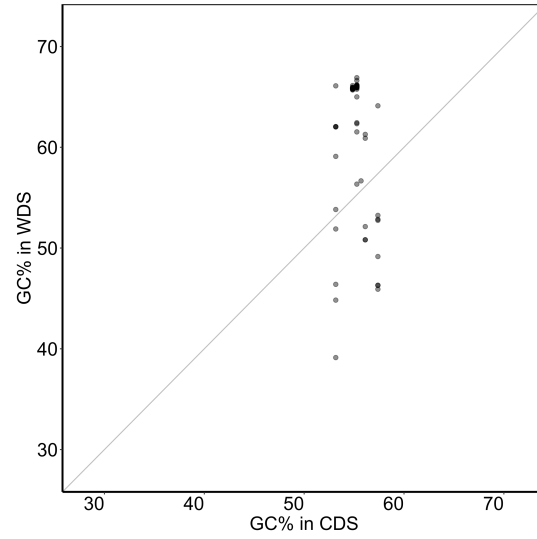


Figure S5.2. The guanine-cytosine (GC) content of DfrB genes identified in a clinical context is not proportional to the GC content of the organism in which they are embedded.
The dataset presented here is from ².

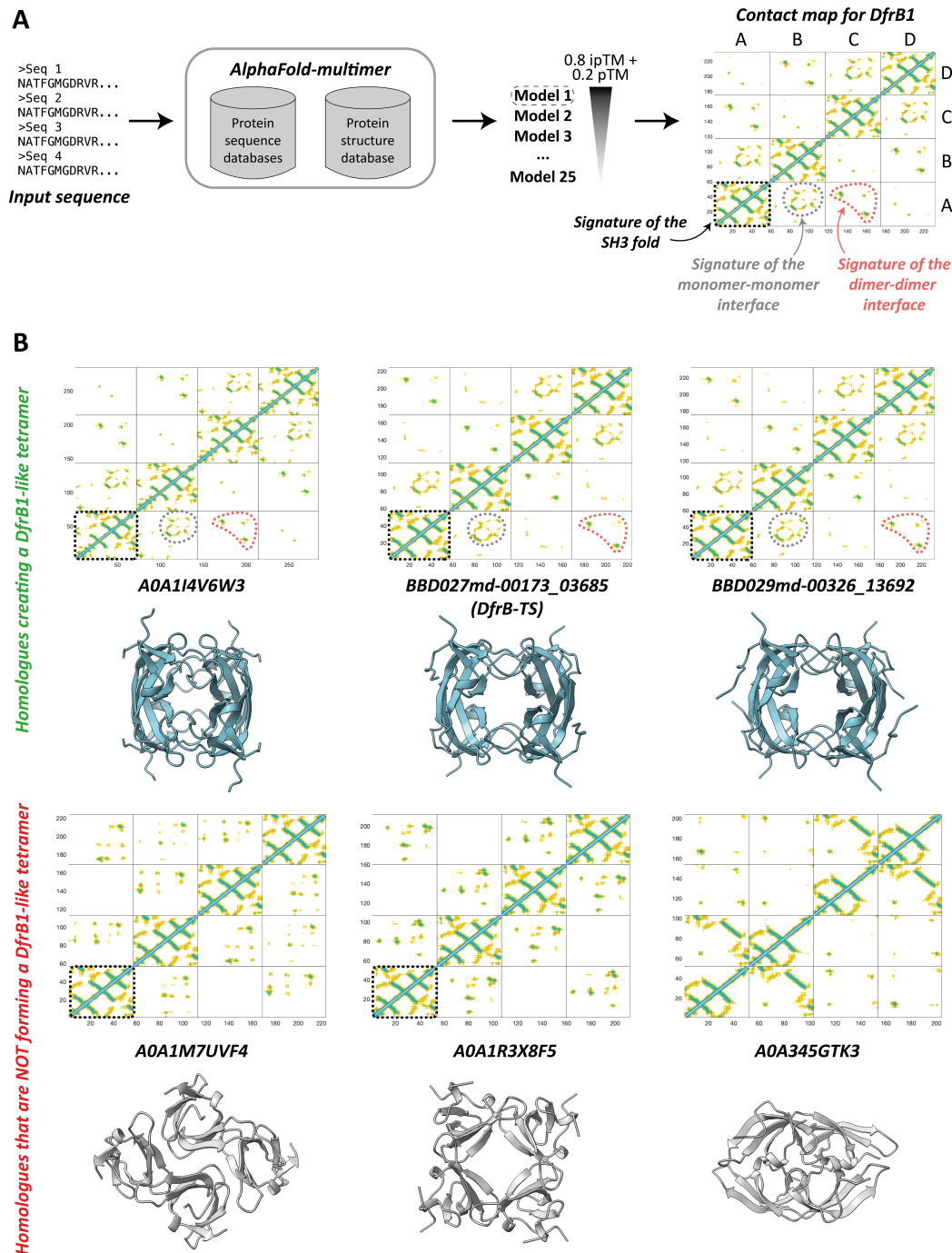


Figure S5.3. Computational pipeline for the analysis of the multimerization prediction for DfrB homologues. **A.** The sequence encoding for the SH3 fold of the DfrB homologues is used as input for AlphaFold-multimer. The algorithm outputs 25 models, which it ranks according to a self-assessment metric. A contact map is produced from the model with the highest value for this metric, which corresponds to the model for which AlphaFold-multimer has the highest confidence. The contact signatures for the SH3 fold, the monomer-monomer interface and the dimer-dimer interface are indicated with dashed lines. **B.** Shown as examples, the contact map of homologues that were characterized in depth in Table 1 are presented, along with their three-dimensional prediction for a homotetrameric complex. The contact signatures for the SH3 fold, the monomer-monomer interface and the dimer-dimer interface are indicated with dashed lines as in **A.**

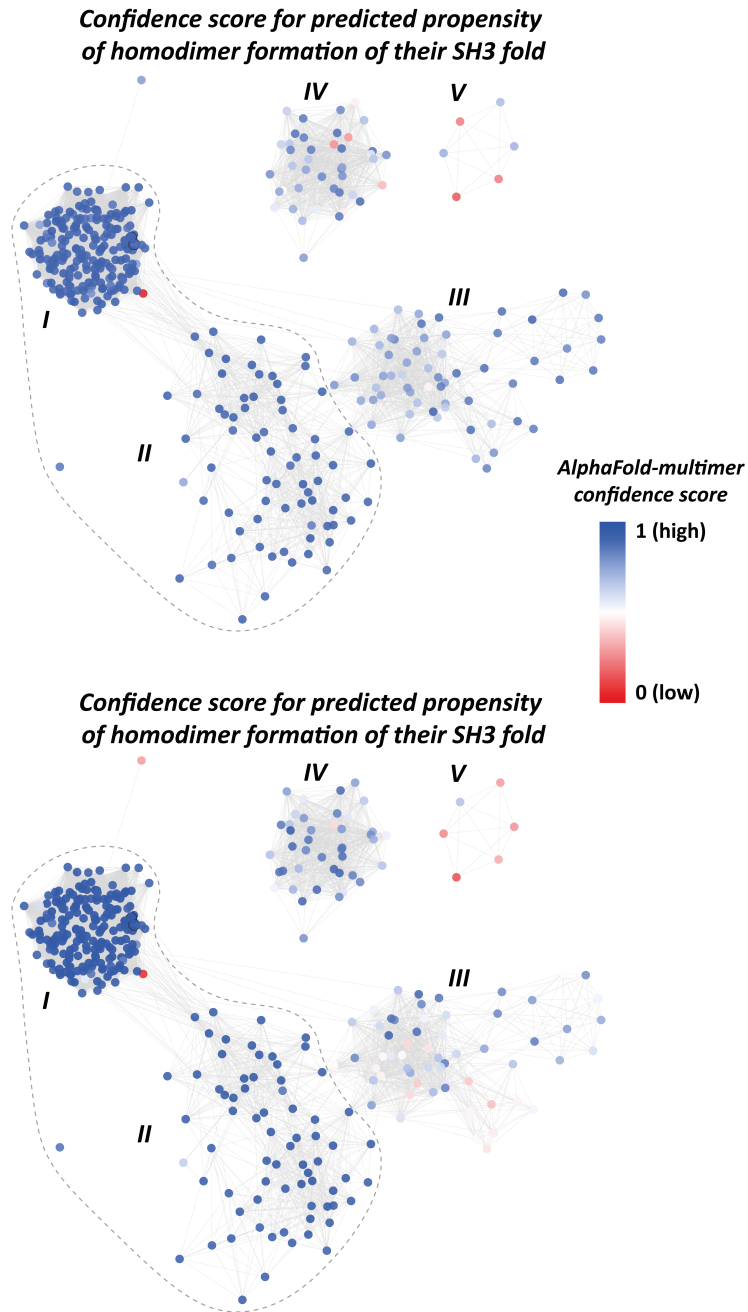


Figure S5.4. AlphaFold-multimer predicts the DfrB1-like complexes in clusters *I* and *II* with high confidence. The confidence score corresponds to 0.8 ipTM (predicted interface TM score) + 0.2 pTM (predicted TM score). The clusters containing DfrB homologues capable of catalyzing the reduction of dihydrofolate are circled with a dotted line.

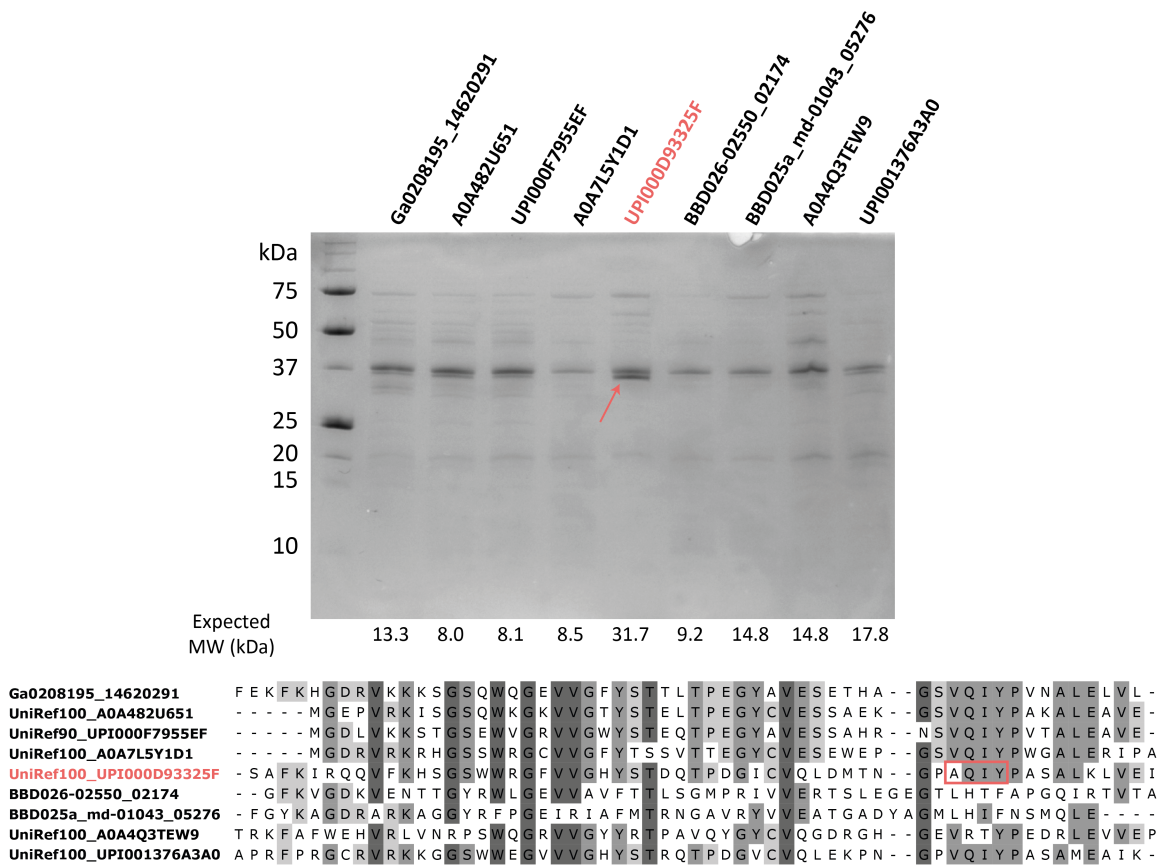


Figure S5.5. The nine DfrB homologues predicted to form a DfrB1-like tetramer but that do not display trimethoprim resistance when expression is induced in *E. coli*.

Top: Tricine SDS-PAGE of the lysate of overexpressed homologues. Only UPI000D93325F displays a visible protein band on gel. Bottom: A MSA of the SH3 domain of these homologues shows that the active site motif of UPI000D93325F (framed in red) has not been observed in other homologues to yield a functional enzyme.

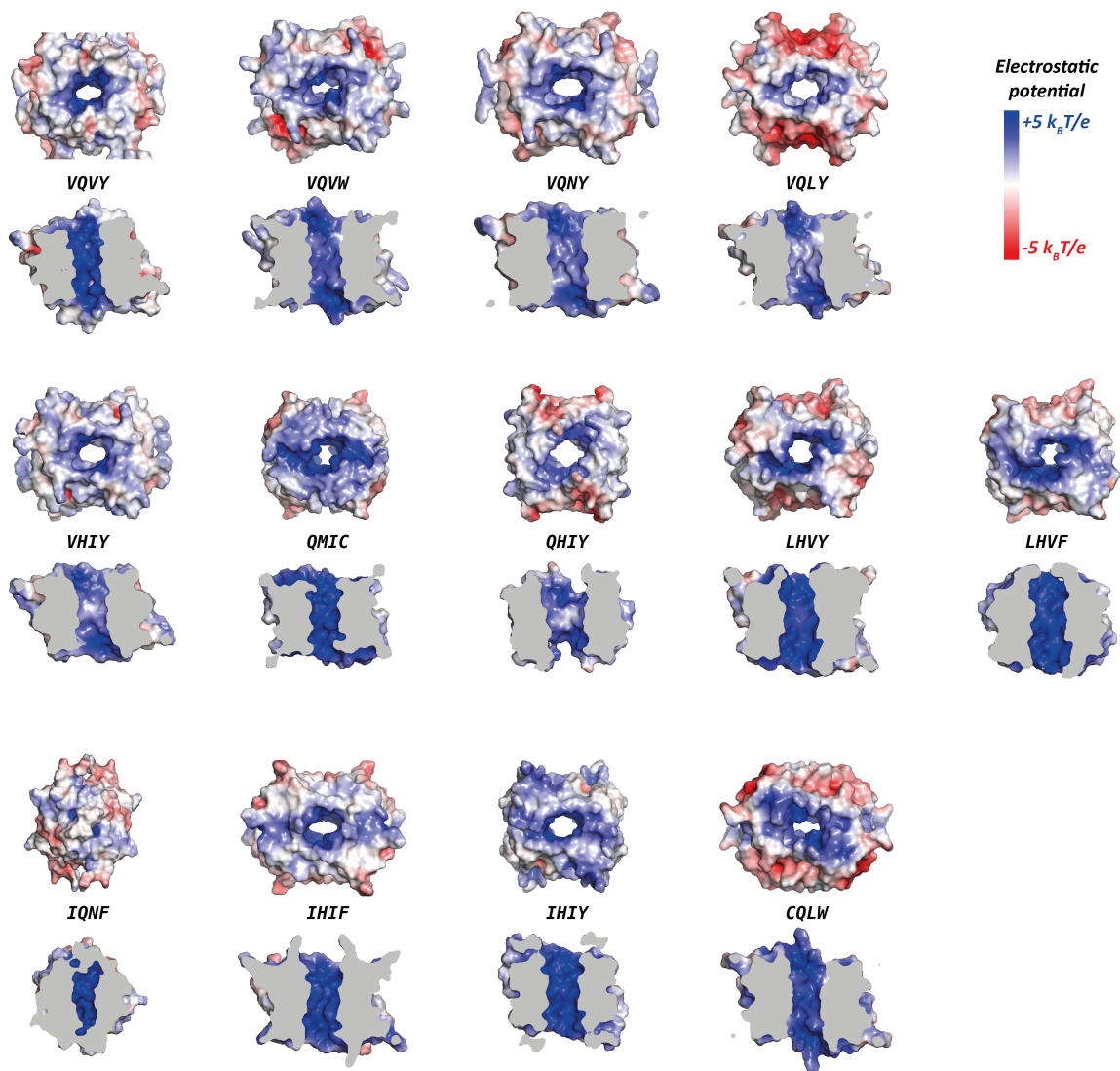


Figure S5.6. The surface electrostatic potential of predicted DfrB1-like tetramers for 13 of the 18 active site motifs.
 The identity of residues on the B4 strand is indicated for each homotetrameric complex.

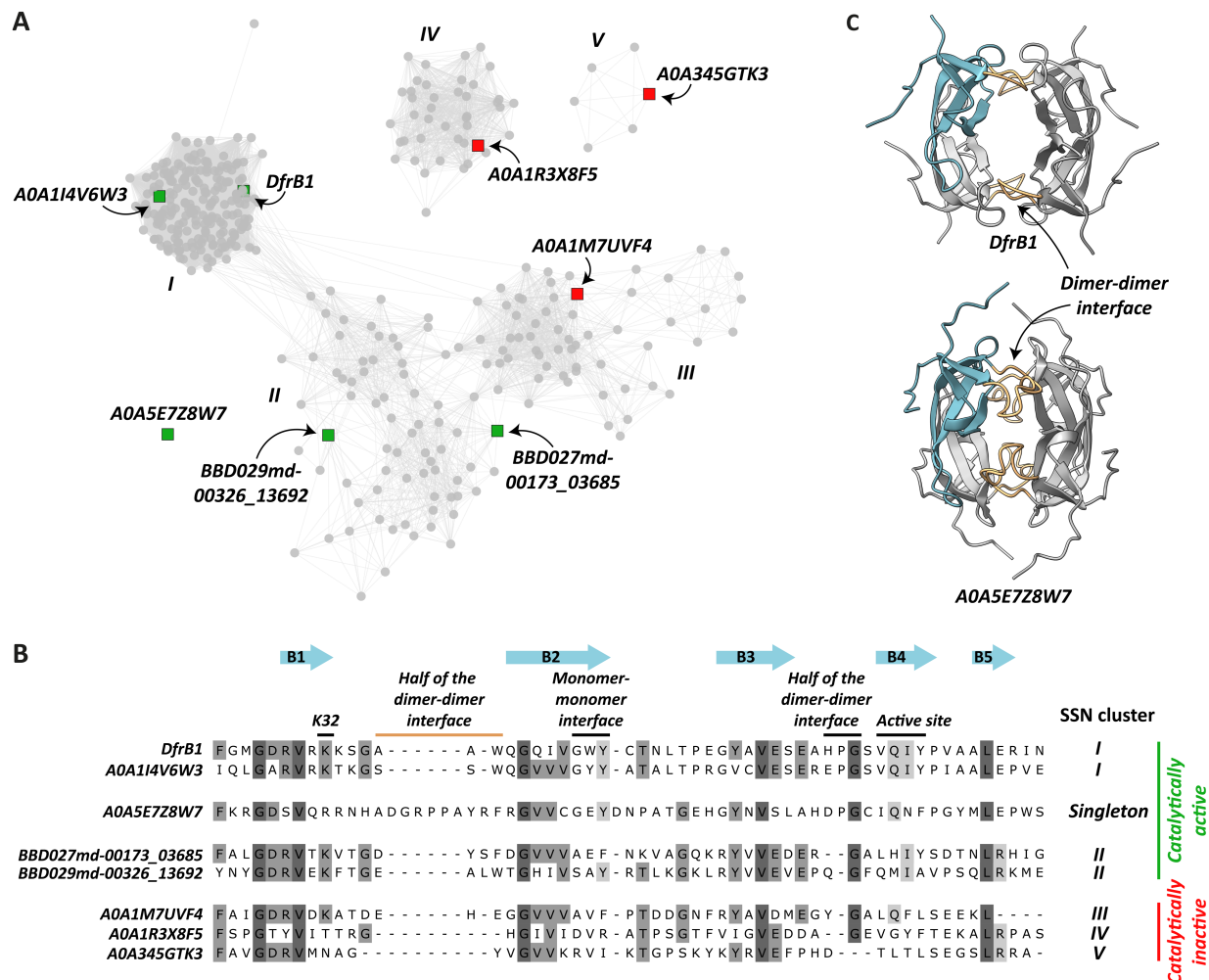


Figure S5.7. DfrB domains across of the DfrB sequence space were characterized *in vitro*.

A. The homologues characterized in depth are identified on the SSN. The representative sequence for DfrB1, DfrB7, is represented. Those colored in green are active as a dihydrofolate reductase, and those colored in red are not. **B.** An alignment of the regions containing the SH3 fold of all characterized homologues performed by MAFFT is presented. Important structural regions of the SH3 fold are indicated. The overall conservation is presented schematically at the top of the MSA, as shown in UGENE³. **C.** The DfrB1 (2rk1 pdb) and the predicted tetramer formed by the SH3 fold of A0A5E7Z8W7 are presented. Only one protomer is colored in blue, and the loop between strand B1 and B2 is colored in orange in both structures.

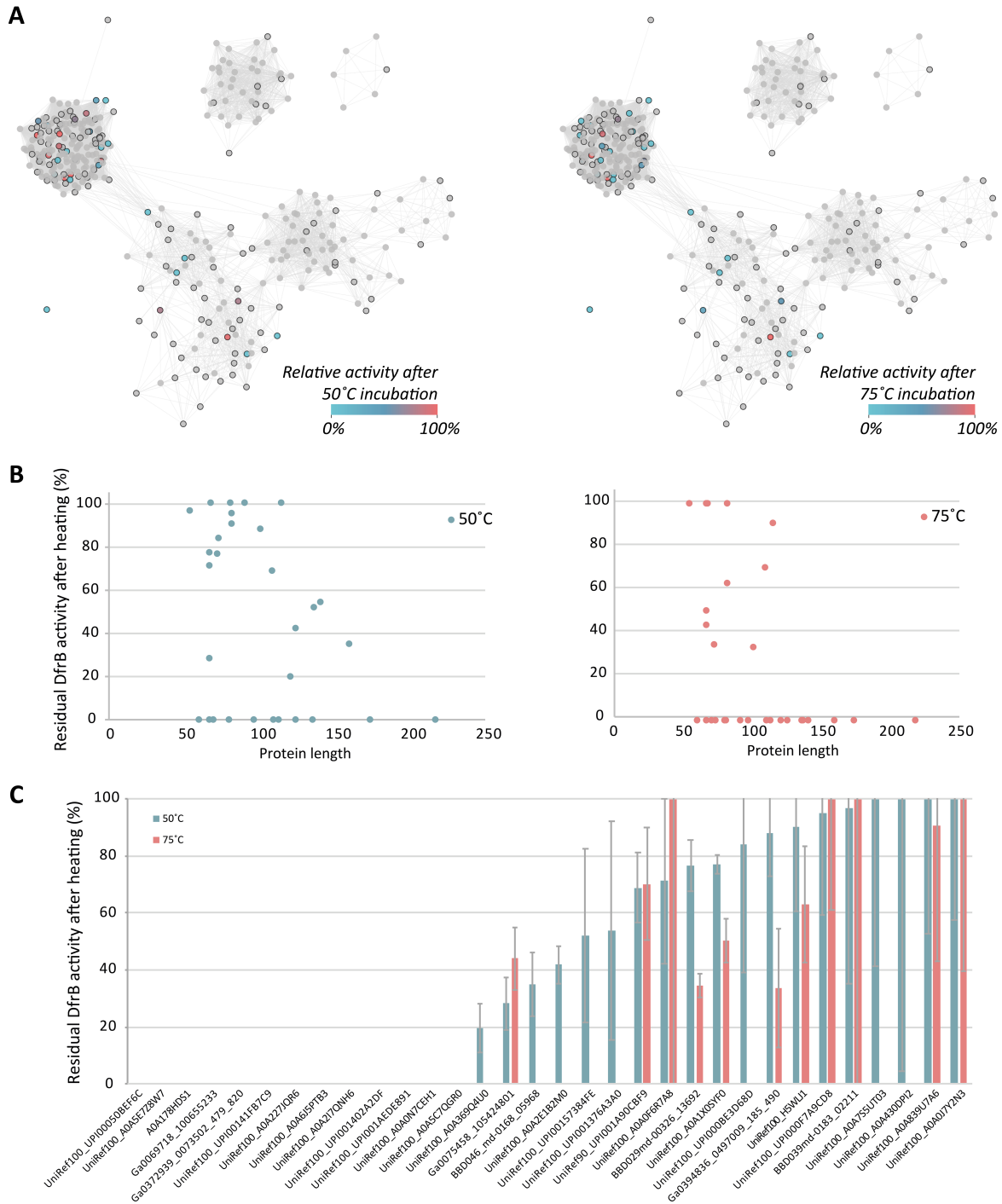


Figure S5.8. Thermotolerance is not a defining property of the DfrB1-like tetramer.

A. In the SSN, homologues with detectable dihydrofolate reductase activity in lysate are color-coded according to their thermostability. Lysates were incubated for 10 minutes at either 50°C or 75°C. All homologues tested for lysate dihydrofolate reductase activity are circled by a black border. **B.** The thermotolerance of each homologue is plotted against its respective sequence length. **C.** Homologues displaying dihydrofolate reductase activity in lysate are identified, and their thermotolerance at 50°C and 75°C is shown. Error bars correspond to the standard deviation.

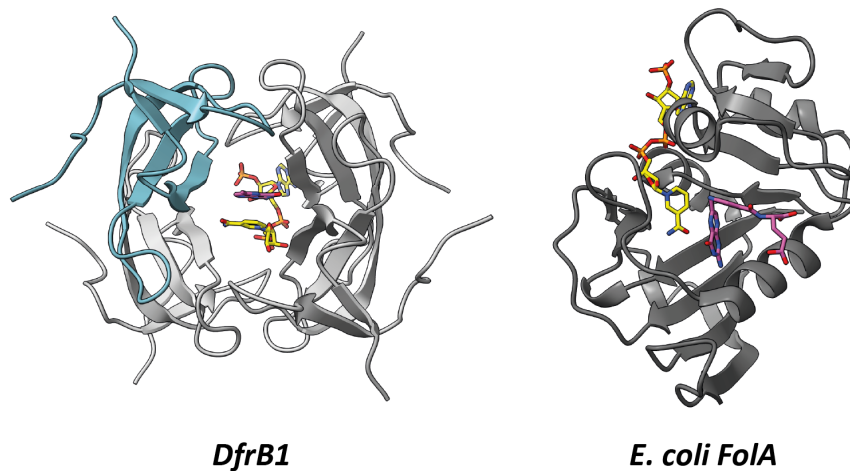
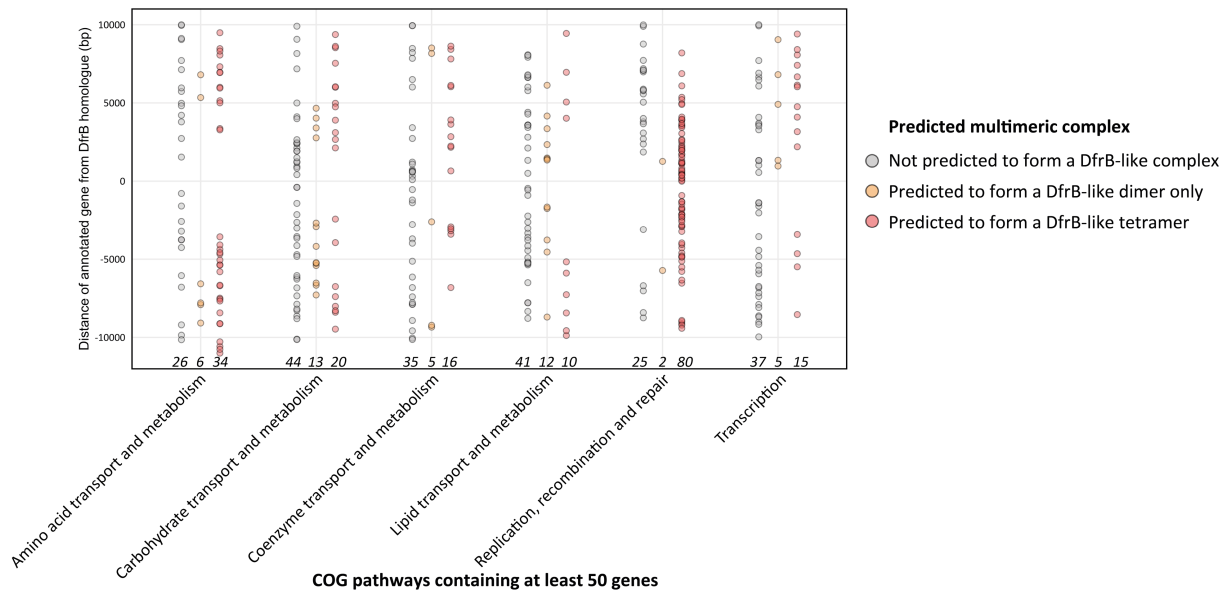


Figure S5.9. The homotetrameric *DfrB1* (pdb 2rk1) and monomeric *E. coli FolA* (pdb 4psy) enzymes both catalyze the reduction of dihydrofolate yet are structurally unrelated. The *DfrB1* homotetramer (one protomer is colored in blue) is presented in complex with NADP⁺ (yellow) and the pterin group of DHF (purple). *E. coli FolA* is in complex with NADP⁺ (yellow) and the DHF analog folate (purple).

A Distance of annotated neighboring genes with DfrB homologues according to their COG pathway



B Conservation of genomic context

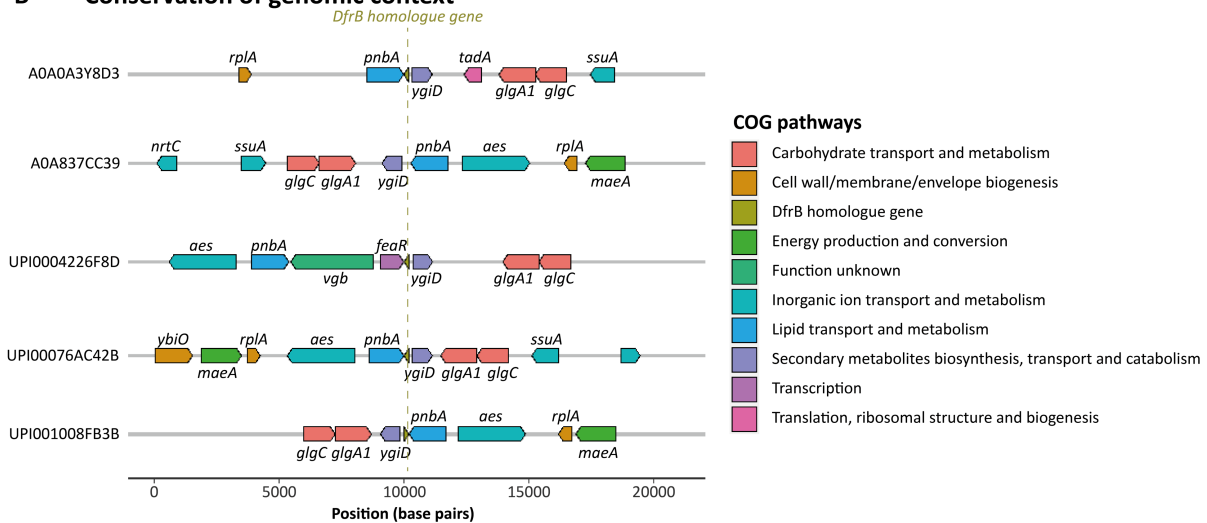


Figure S5.10. Genomic context of DfrB homologues.

A. Neighboring genes associated with a COG pathway are mapped according to their distance from the *dfrB* homologues. Only COG pathways containing at least 50 genes are shown. Genes are binned according to the predicted multimeric assembly of the DfrB homologue in the same genomic context. The number of genes in each category is shown at the bottom of the graph. **B.** The conserved genomic context of five *dfrB* homologues predicted to form a DfrB1-like dimer, but not a tetramer, is mapped. These homologues share between 76% and 90% local sequence identity.

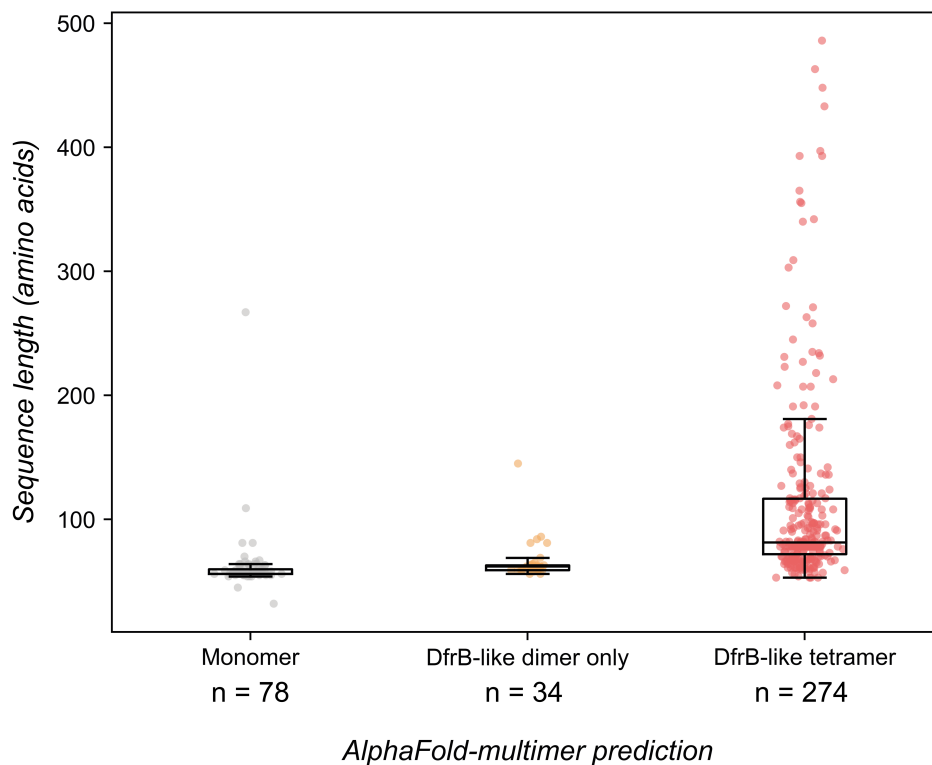


Figure S5.11. Relationship between sequence length and acquisition of the tetramerization property in DfrB homologues.

Each data point represents the sequence length of a DfrB homologue, and boxes extend from the first quartile to the third quartile of the data, with a line at the median and whiskers extending to the farthest data point lying within $1.5 \times$ the interquartile range from the box. Homologues are categorized into three groups based on AlphaFold-multimer predictions: 'Tetramer' for those predicted to form a DfrB1-like tetramer, 'Dimer' for those predicted to form only a DfrB1-like dimer, and 'Monomer' for those not predicted to resemble the DfrB complex. The data reveal a significant correlation between increased sequence length and the acquisition of the ability to form a DfrB1-like tetramer (p-value = 9.6×10^{-36} , Mann–Whitney U test, 'Tetramer' vs. combined 'Dimer' and 'Monomer' categories).

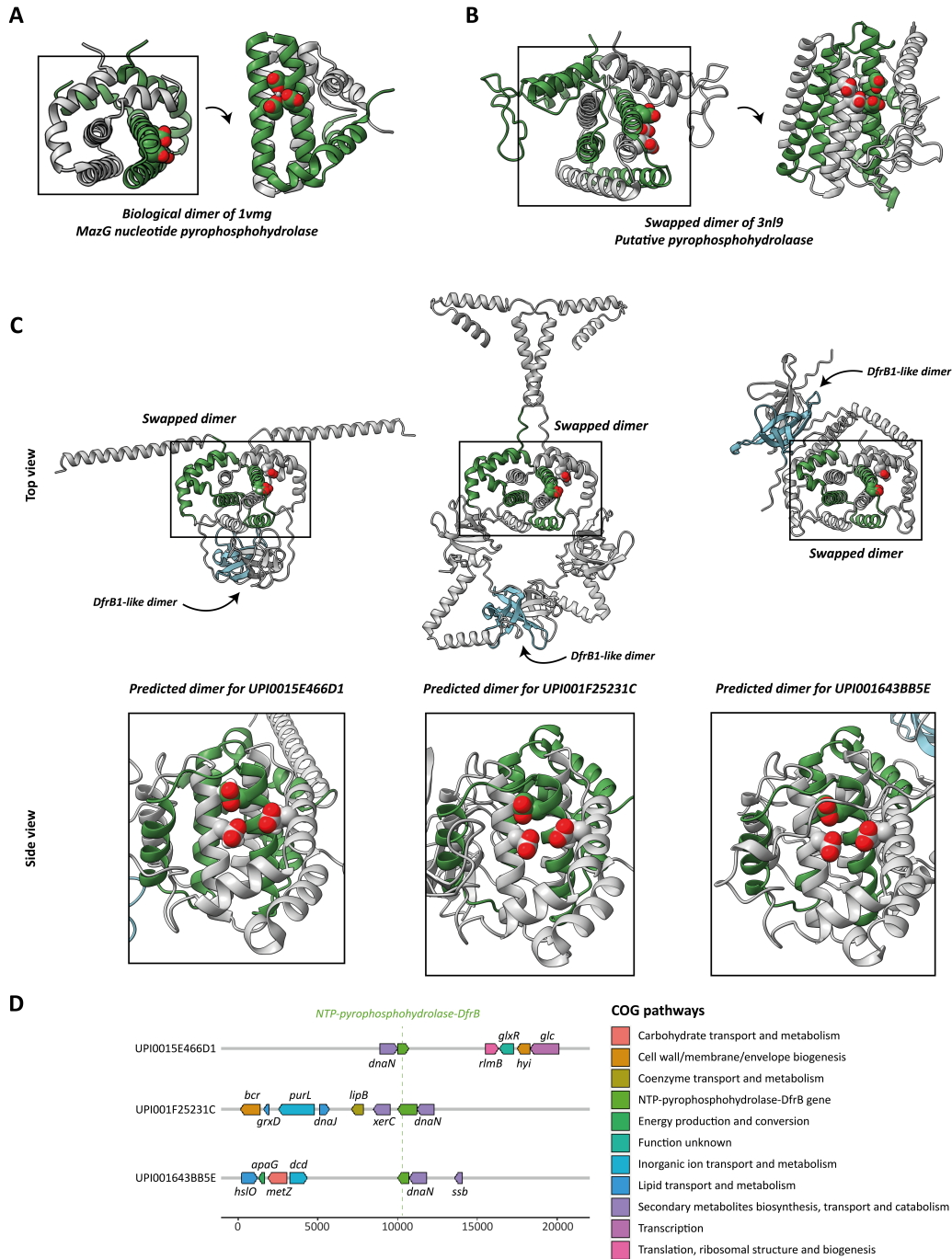


Figure S5.12. NTP pyrophosphohydrolase domains fused to DfrB domains.

A. The canonical dimer of the NTP pyrophosphohydrolase domain is shown, from PDB 1vmg. The side chains of negatively charged residues known to bind a magnesium atom are represented on one of two symmetric sides. **B.** The swapped dimer of the NTP pyrophosphohydrolase domain is shown, from PDB 3n19. The conserved side chains of negatively charged residues known to bind a magnesium atom are represented on one of the two symmetric sides. **C.** The predicted dimers of the three NTP pyrophosphohydrolase-DfrB fusions are represented. The DfrB domain of one protomer is colored in blue, and the NTP pyrophosphohydrolase domain is colored in green. The side chains of conserved residues known to bind a magnesium atom are represented on one of the two symmetric sides. **D.** The genomic context of all three fusion genes is annotated and colored according to the COG pathway with which each gene is associated.

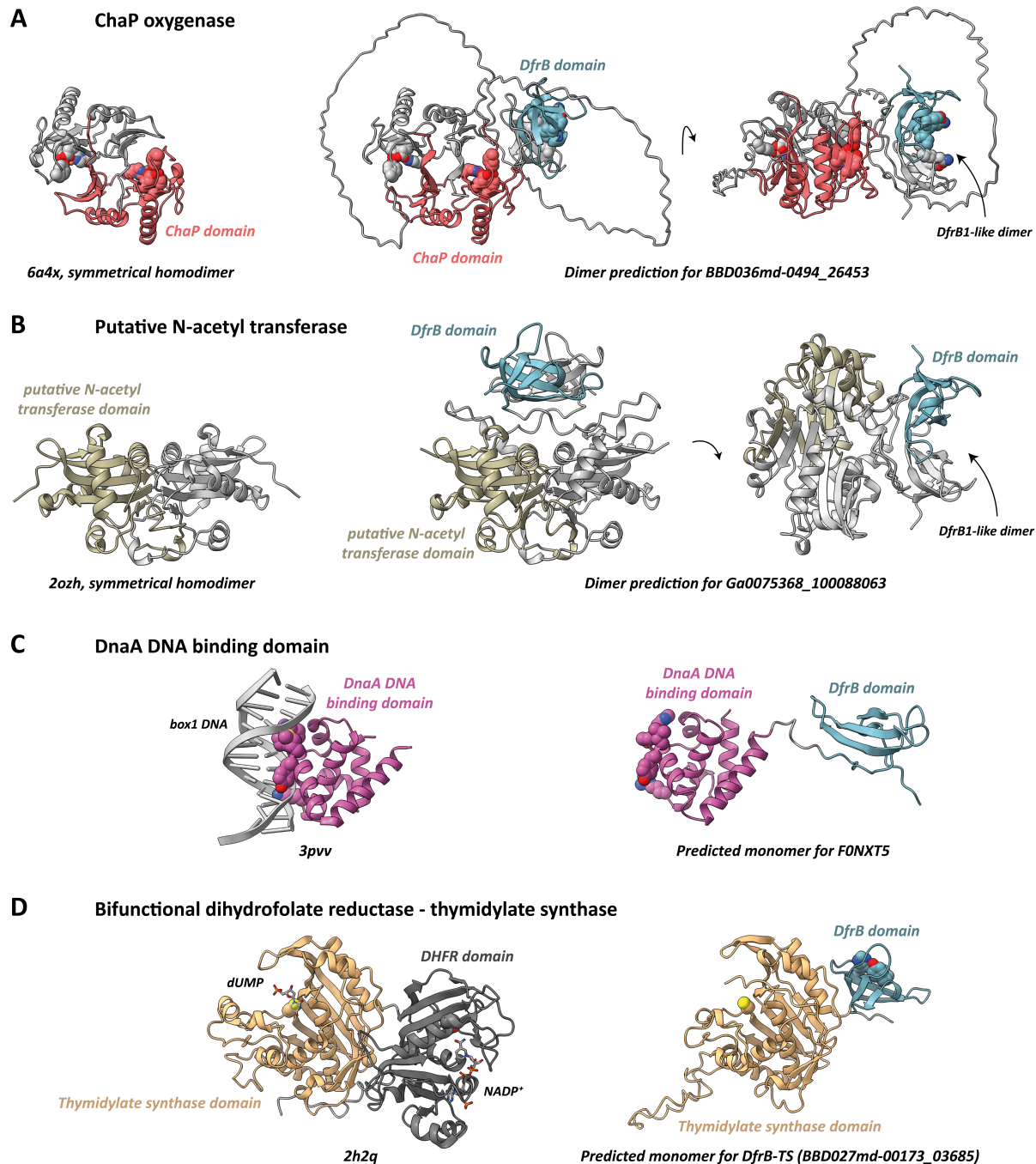


Figure S5.13. The DfrB domain is fused to a variety of characterized domains.

A. The ChaP dioxygenase domain, known to form a symmetrical dimer, is fused to a DfrB domain. The ChaP oxygenase domain of BBD036md-0494_26453 shares 39% sequence identity and a TM-score of 0.92 with the closest PDB template 6a4x. The residues from 6a4x that form a complex with iron (H63, Y109, E119, and Y125), represented as spheres, are conserved in BBD036md-0494_26453, along with the VQIY active site motif present in the DfrB domain. The dimer prediction exhibits the expected topological arrangement for both the ChaP and DfrB domains. **B.** A putative *N*-acetyl transferase domain is fused to a DfrB domain. The *N*-acetyl transferase domain of Ga0075368_100088063 shares 49% sequence identity and a TM-score of 0.97318 with the closest PDB template, 2ozh. The dimer prediction of this protein showcases the characteristic topology of both the *N*-acetyl transferase and DfrB domains. **C.** The DnaA DNA binding domain, as crystallized in complex with the origin of replication *oriC*

(PDB 3pvv), is found in association with a DfrB domain in F0NXT5. Key residues responsible for DNA binding, with their side chains represented as spheres, are mostly conserved in F0NXT5, demonstrating the importance of this complex in DNA-related processes (3pvv- F0NXT5 residues identity: K436R, D469D, H470H, T471A, M474L, and Y475Y). **D.** The canonical DHFR-thymidylate synthase (TS) fusion, illustrated in complex with dUMP and NADP⁺ (in sticks), is mirrored by the structure of BBD027md-00173_03685, revealing the fusion of TS with the DfrB domain. The figure represents in spheres the side chains of the TS domain catalytic serine, the canonical DHFR active site aspartate (left), and the active site residues of the DfrB domain (right), featuring the characteristic LHIY motif.

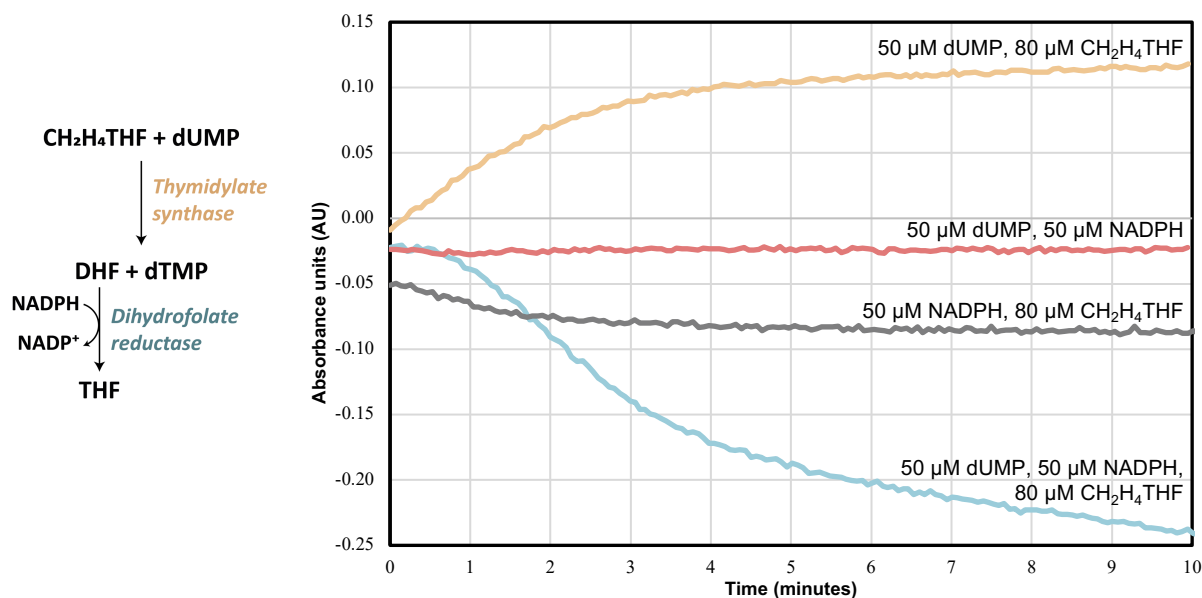


Figure S5.14. The DfrB-TS enzyme is bifunctional, acting as both a thymidylate synthase (TS domain) and a dihydrofolate reductase (DfrB domain).

In beige: the increase in absorbance signal over time when CH₂H₄THF and dUMP are incubated with DfrB-TS reflects the production of dihydrofolate (DHF) as a result of TS domain activity. In blue: when NADPH is added to both substrates of the TS domain, the DHF produced by the TS domain is reduced by the DfrB domain, resulting in a decrease in the absorbance signal over time; the lag is consistent with the need to build up DHF product of the TS reaction to allow the DfrB reaction to reach a significant rate. Combining either reagent of the TS reaction with NADPH (red and grey curves) gives no reaction. The average of three replicates is shown.

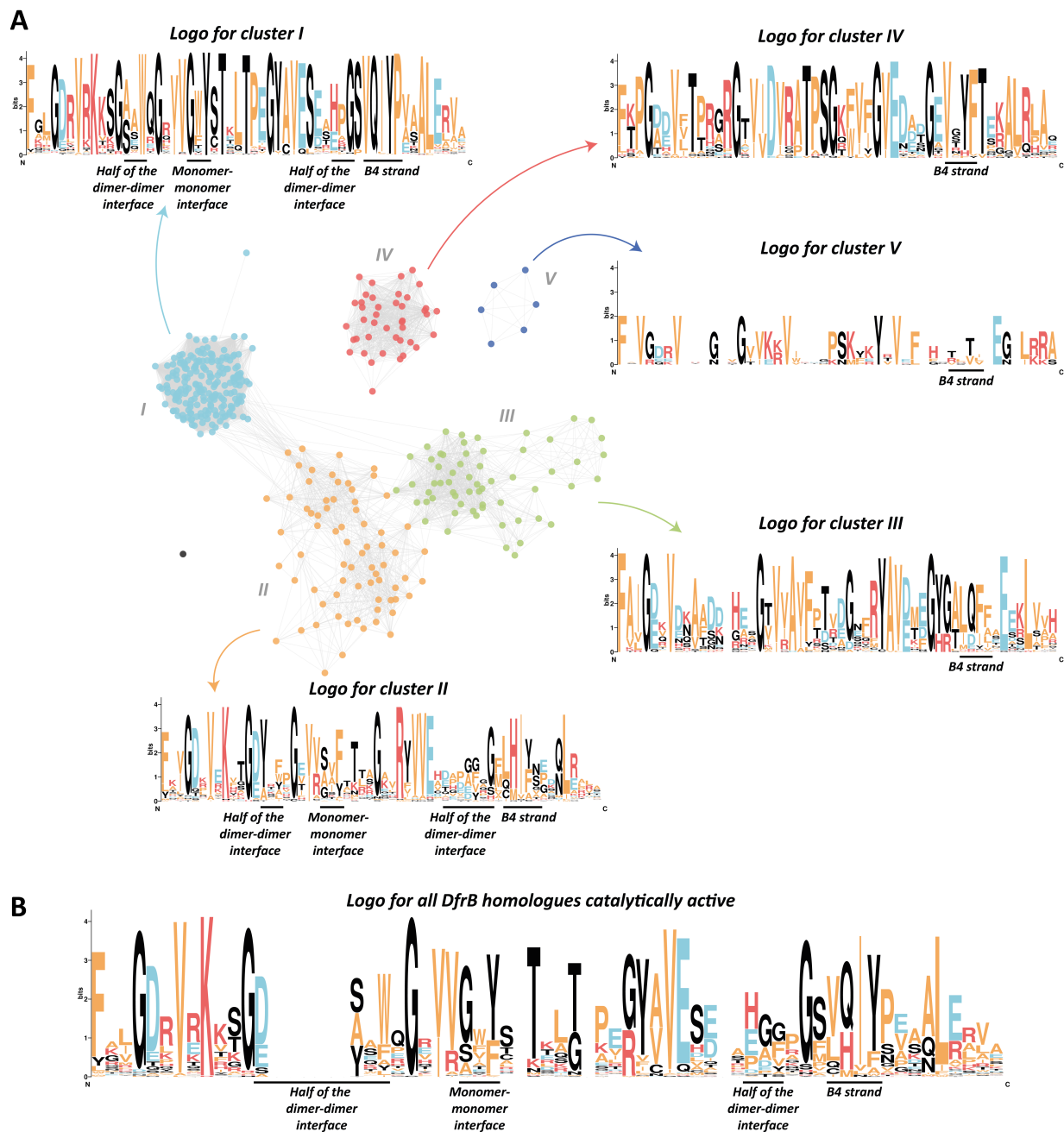


Figure S5.15. Sequence logos are shown for each cluster of the DfrB sequence space.

A. Alignment of the SH3 fold of homologues from the five clusters have been performed by MAFFT, gaps were removed, and the logo were performed by WebLogo⁴. Structurally important elements are annotated. **B.** The logo created from the alignment of the SH3 fold of all sequences experimentally confirmed to provide trimethoprim resistance when overexpressed in bacteria is presented. Gaps within the alignment were not removed.

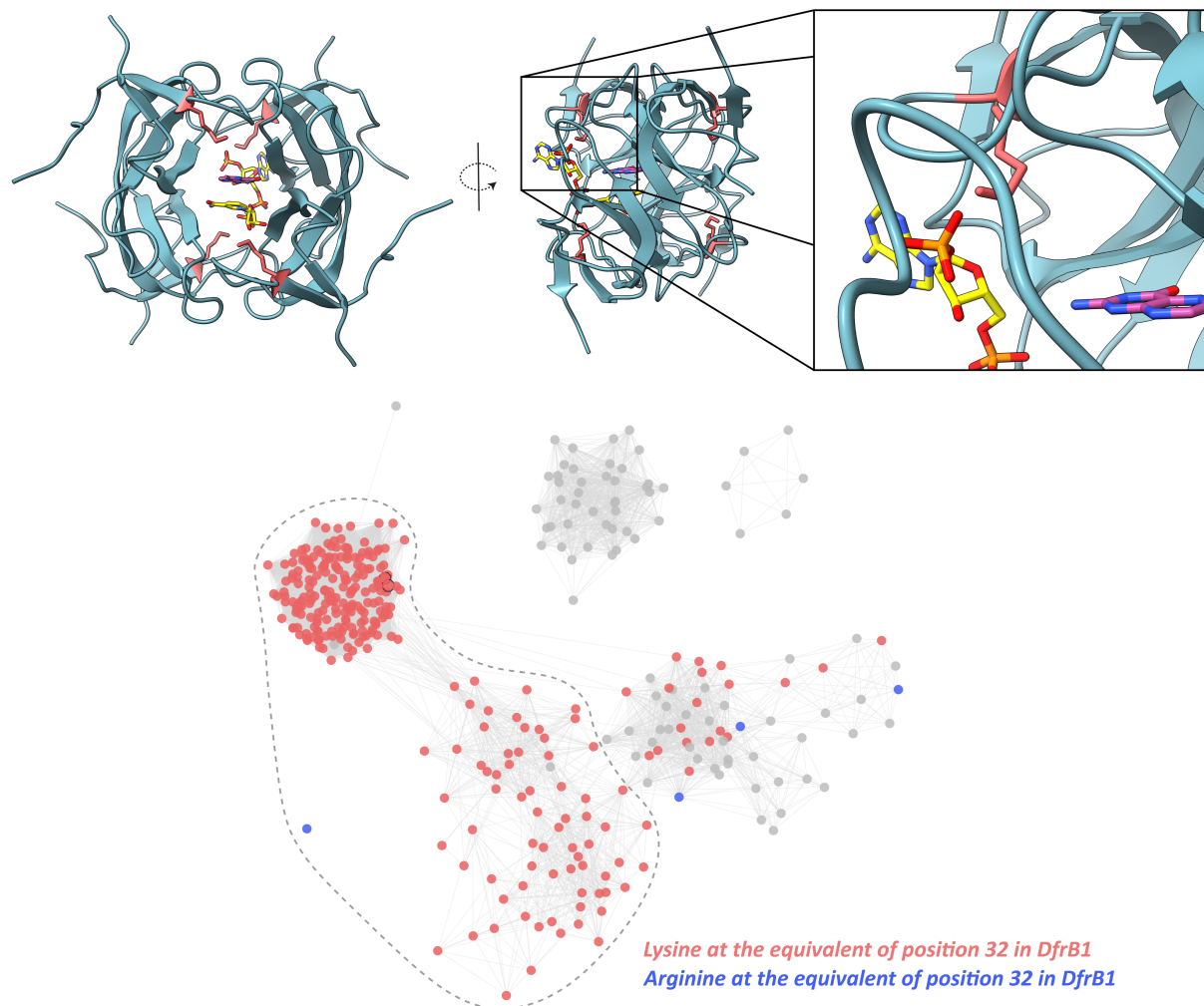


Figure S5.16. Lys32 is functionally important for the DfrB domain.

Top: Lys32 is represented by red sticks on the tertiary complex of DfrB1 (pdb 2rk1). The NADP⁺ molecule is colored in yellow, and the pterin group of DHF is colored in purple. The side chain of one of the Lys32 is shown to interact with one of the phosphate groups of NADP⁺ in the crystal structure of the complex. Bottom: the SSN is colored according to the identity of the residue at position 32. The clusters where DfrB homologues are capable of catalyzing the reduction of dihydrofolate are circled by a dotted line.

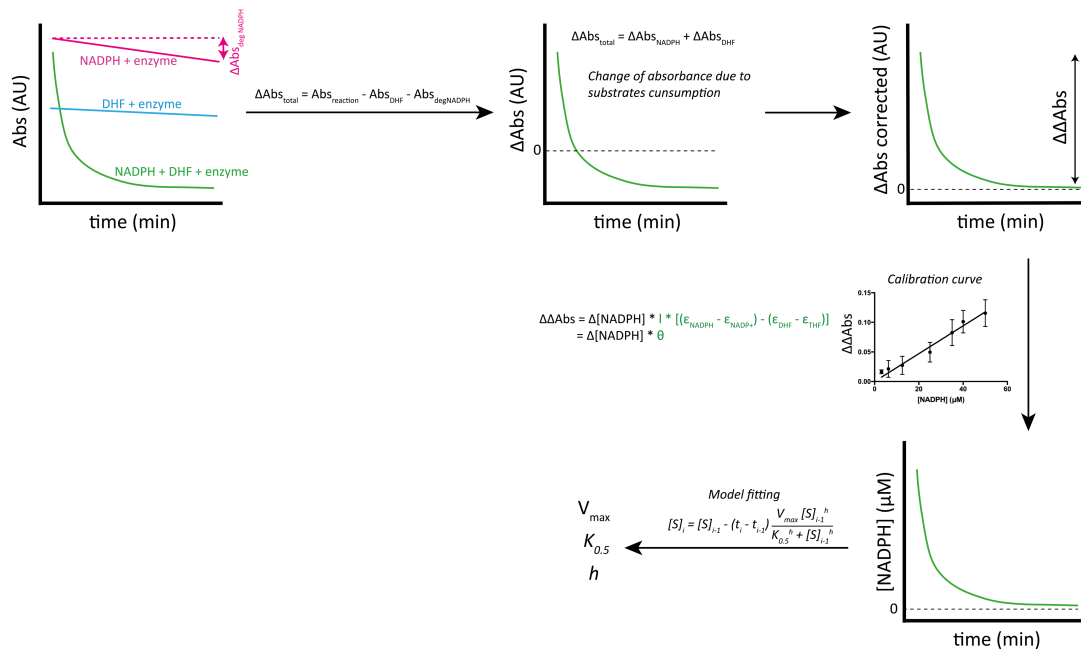


Figure S5.17. Pipeline for the calculation of kinetic parameters using data from a single activity curve.

The difference in absorbance when the clarified lysate of a DfrB homologue is incubated with excess DHF and a limiting concentration of NADPH is measured until complete consumption of NADPH. The calibration curve for NADPH was prepared as follows: The degradation of the substrates, as well as the absorbance of the excess DHF, are subtracted from the absorbance measurements. The total difference in absorbance measurements ($\Delta\Delta\text{Abs}$) for different starting concentration of NADPH, always in large excess of DHF, is used to construct the calibration curve. The calibration curve allows translating ΔAbs measurements into concentration of NADPH. The concentration of NADPH in relation to time is fitted to a kinetic model allowing extraction of the maximum velocity (V_{\max}), half maximum concentration constant ($K_{0.5}$) and the Hill coefficient (h). This pipeline was inspired by⁵.

5.12 Supplementary references

- (1) Lemay-St-Denis, C.; Pelletier, J. N. From a Binding Module to Essential Catalytic Activity: How Nature Stumbled on a Good Thing. *Chem. Commun.* **2023**, 59 (84), 12560–12572. <https://doi.org/10.1039/D3CC04209J>.
- (2) Lemay-St-Denis, C.; Diwan, S.-S.; Pelletier, J. N. The Bacterial Genomic Context of Highly Trimethoprim-Resistant DfrB Dihydrofolate Reductases Highlights an Emerging Threat to Public Health. *Antibiotics* **2021**, 10 (4), 433. <https://doi.org/10.3390/antibiotics10040433>.
- (3) Okonechnikov, K.; Golosova, O.; Fursov, M. Unipro UGENE: A Unified Bioinformatics Toolkit. *Bioinformatics* **2012**, 28 (8), 1166–1167. <https://doi.org/10.1093/bioinformatics/bts091>.
- (4) Crooks, G. E.; Hon, G.; Chandonia, J.-M.; Brenner, S. E. WebLogo: A Sequence Logo Generator: Figure 1. *Genome Res.* **2004**, 14 (6), 1188–1190. <https://doi.org/10.1101/gr.849004>.
- (5) Tamer, Y. T.; Gaszek, I. K.; Abdizadeh, H.; Batur, T. A.; Reynolds, K. A.; Atilgan, A. R.; Atilgan, C.; Toprak, E. High-Order Epistasis in Catalytic Power of Dihydrofolate Reductase Gives Rise to a Rugged Fitness Landscape in the Presence of Trimethoprim Selection. *Mol Biol Evol* **2019**, 36 (7), 1533–1550. <https://doi.org/10.1093/molbev/msz086>.

Chapitre 6. Discussion

6.1 L'étude des mécanismes à la base de l'évolution d'enzymes ouvre des horizons

C'est en analysant les enzymes qui ont été naturellement explorées durant l'évolution qu'on peut répondre à des questions fondamentales biochimiques. Parmi celles-ci, on retrouve : « quels sont les déterminants de la fonction d'une enzyme ? » ou encore « comment une fonction catalytique peut émerger d'un domaine protéique ? ».

Comme présenté au Chapitre 1, les technologies développées au cours des dernières décennies ont donné les moyens de répondre à ces ambitieuses questions. Dans cette thèse, j'ai mis à profit ces nouvelles technologies pour étudier l'évolution des enzymes DfrB. En particulier, les résultats présentés dans cette thèse ont informé sur deux volets de l'évolution des DfrB : i) leur évolution et la distribution récente des gènes *dfrB* et ii) les propriétés qui ont mené à l'évolution d'une capacité catalytique chez le domaine DfrB.

6.2 L'évolution récente des gènes *dfrB*

6.2.1 L'identification des *dfrB*

Pour retracer l'évolution des DfrB, il est essentiel d'obtenir une quantité critique de données de séquences. Pourtant, comme discuté au Chapitre 3, les gènes des DfrB1 à DfrB9 étaient considérés comme marginaux jusqu'en 2019, puisqu'ils étaient rarement répertoriés. Plusieurs explications peuvent être avancées à cet égard. Tout d'abord, leur identification était, et est encore souvent¹, réalisée par PCR (*Polymerase Chain Reaction*). Cette méthode de recherche de gènes d'intérêt couramment appliquée dans le contexte de la résistance aux antibiotiques utilise une banque d'amorces spécifiques qui exclut généralement les DfrB, justement puisqu'on les considère comme peu courants. En d'autres termes, on ne peut pas les identifier si on les exclut de la recherche.

De plus, la taille des gènes des DfrB (237 paires de bases), leur utilisation en codons et leur contenu en GC (Figure S5.2) diffèrent de ceux des autres gènes présents dans les génomes d'organismes associés à la résistance aux antibiotiques, soit l'unique contexte dans lequel on avait retrouvé les DfrB jusqu'en 2019. Par conséquent, les méthodes de prédiction de gènes *ab initio* ne parviennent pas systématiquement à identifier les *dfrB* ; il arrive parfois que ces algorithmes annotent plutôt un gène alternatif, et que le gène *dfrB* ne soit pas identifié. Ainsi, pour détecter spécifiquement la présence de DfrB, il est préférable d'utiliser des méthodes de prédiction basées sur la similarité, en se référant à une base de données contenant les gènes *dfrB* validés expérimentalement. Pourtant, il n'est pas possible, avec l'information disponible jusqu'à

présent, de connaître l'efficacité de ces méthodes par similarité à détecter toutes les DfrB présentes dans les séquences génomiques et métagénomiques. Pour y remédier, il serait possible de comparer tous les cadres de lecture possibles aux gènes identifiés par diverses méthodes prédictives dans une grande variété de génomes, et d'évaluer leur efficacité.

Enfin, l'augmentation récente des analyses métagénomiques d'échantillons environnementaux a révélé que la rareté des DfrB peut s'expliquer par le type d'organismes séquencés dans les bases de données publiquement disponibles. Au Chapitre 3, j'ai présenté nos efforts datant de 2021, visant à explorer de manière plus globale la distribution des DfrB au sein de bases de données génomiques publiquement disponibles. Bien que cela nous ait permis d'identifier, pour la première fois, des DfrB dans des contextes autres que la résistance aux antibiotiques, cette découverte représentait une proportion mineure de notre jeu de données. Ceci est dû au type de données que ces bases de données comprennent, qui sont largement peuplées par des données cliniques. Par exemple, la proportion de paires de bases provenant d'échantillons environnementaux au sein de la base de données GenBank s'élevait à uniquement 0,27% en 2016.²

6.2.2 La métagénomique pour étudier l'évolution des DfrB

En 2022, le laboratoire Pelletier a publié une étude, dirigée par Stella Cellier-Goetghebeur, qui a identifié dix nouveaux gènes *dfrB* dans des échantillons métagénomiques environnementaux.³ Ces gènes ont été expérimentalement validés pour leur capacité à conférer une résistance au triméthoprim lorsqu'ils sont surexprimés dans des bactéries. Dans la même année, David Kneis et ses collègues ont publié une étude, également avec une approche métagénomique, où des gènes *dfrB* ont été identifiés dans des échantillons environnementaux de rivières non contaminées.⁴ En 2023, une collaboration entre l'équipe de David Kneis et le laboratoire Pelletier a mené à la publication d'une étude où l'on a quantifié l'abondance des gènes *dfrB* dans des environnements aquatiques séquencés de manière métagénomique.⁵ Cette étude a révélé que les communautés bactériennes aquatiques constituent un réservoir important de certains gènes *dfrB* ; d'intérêt particulier, ces travaux ont révélé que les gènes *dfrB* identifiés dans les environnements aquatiques liés aux activités humaines – telles les eaux usées – comportent des gènes *dfrB* distincts de ceux identifiés en environnements aquatiques non touchés par les activités humaines. Enfin, une étude métagénomique datant de 2021, menée par Gonçalo Macedo, a révélé une abondance significative du gène DfrB3 dans des échantillons de sol non contaminé par les activités humaines.⁶

Ce cas de figure, où une famille de protéines se révèle plus abondante dans des échantillons métagénomiques que dans les séquences génomiques, est commun. En 2023, Pavlopoulos et ses collègues ont révélé l'énorme espace de séquences présents dans les échantillons environnementaux.⁷ Ils ont identifié plus de 100 000 familles de protéines se situant exclusivement dans des échantillons métagénomiques, alors qu'un peu plus de 90 000 familles sont compilées à partir des génomes d'organismes de référence. Cette

étude a révélé l'aspect instrumental que représentent des données métagénomiques pour l'exploration de l'espace de séquences, qui s'est révélé tout aussi essentiel à l'étude de l'évolution des DfrB dans le cadre de cette thèse.

Les études métagénomiques rapportant des DfrB ont également mis en évidence une tendance inhabituelle. La modification de la composition des milieux naturels par l'introduction de matières reliées aux activités humaines semble réduire l'abondance des gènes *dfrB* dans les échantillons prélevés, ce qui contraste avec ce qui est généralement observé pour d'autres types de gènes de résistance aux antibiotiques. En effet, typiquement, les gènes bactériens de résistance sont absents d'échantillons environnementaux non contaminés ; leur dissémination dans l'environnement est causée par le rejet de déchets découlant de l'activité humaine.⁸ L'observation de cette tendance inusitée été faite à la fois dans l'étude menée par Gonçalo Macedo, où l'introduction de déchets animaux comme engrais a réduit l'abondance du gène *dfrB3* dans le sol,⁶ et dans l'étude dirigée par David Kneis, où l'introduction d'eaux usées municipales dans une rivière a réduit significativement l'abondance des gènes *dfrB*.⁴ Ces découvertes suggèrent que les gènes *dfrB* sont présents de manière endogène dans des organismes environnementaux, dont les populations sont diminuées significativement par l'introduction de déchets, qui comprennent des organismes bactériens liés aux activités humaines.

6.2.3 L'évolution génomique des DfrB

En somme, l'identification de DfrB au sein de plusieurs bases de données métagénomiques et de plusieurs types d'environnements non associés aux activités humaines réfute l'idée que ces protéines aient évolué leur capacité à réduire le dihydrofolate en réponse à l'introduction clinique du triméthoprim qui inhibe les dihydrofolate réductase microbiennes de type F_{olA}. En effet, des gènes partageant au moins 95% d'identité de séquences avec les *dfrB* sont présents naturellement dans des environnements non exposés aux activités humaines. La pression de sélection due à l'introduction du triméthoprim a probablement conduit à une évolution génomique plutôt qu'à une évolution moléculaire de ces protéines. Il est probable que les gènes *dfrB* aient été acquis par des bactéries pathogènes via des transferts horizontaux de gènes, leur procurant la capacité de proliférer en présence de triméthoprim. Par contre, la mécanistique génomique ayant mené à l'intégration des DfrB dans la clinique est encore inconnue.

Quatre étapes principales ont été rapportées pour qu'un gène chromosomal et immobile d'un organisme environnemental émerge au sein d'un pathogène et participe au problème de la résistance aux antibiotiques.⁹ D'abord, le gène doit acquérir la capacité à se déplacer au sein d'un génome, par l'acquisition de séquences d'insertion ou encore par son insertion dans un intégron. Ensuite, le gène doit être relocalisé au sein d'un élément génétique permettant de se déplacer de manière autonome entre les cellules, tel qu'un plasmide ou un élément de conjugaison. Puis, c'est le transfert horizontal du gène dans un organisme pathogène, qui

peut se faire via plusieurs organismes intermédiaires. La quatrième étape est le transfert physique de l'organisme portant le gène au sein d'un microbiote humain ou animal.

Pour récapituler ces étapes dans le cas de la DfrB, il sera essentiel d'avoir accès à des données de séquences de milieux environnementaux – le rhizobium et les milieux aquatiques de rivières étant des réservoirs importants – et de divers milieux ayant été exposés au triméthoprime. Alors que la métagénomique constitue l'approche clé à cette analyse puisqu'elle permet de séquencer les échantillons non cultivés provenant d'une grande variété d'organismes, les séquences de *contigs* très courtes qu'elle peut générer limite souvent significativement les analyses qu'on peut en faire. Puisque la taxonomie et la présence d'éléments génomiques sont essentielles à l'analyse de leur évolution récente, les séquences récoltées devront avoir une longueur suffisante à l'identification de ces éléments.

Si cette analyse est fructueuse et que l'émergence des DfrB dans le milieu clinique peut être récapitulée, on pourra répondre à plusieurs interrogations latentes, tel que i) La diversité de gènes *dfrB* dans le résistome peut-elle être attribuée à plusieurs événements de transfert de gènes *dfrB* distincts entre des organismes environnementaux et des pathogènes, ou est-elle le résultat d'une évolution moléculaire suivant l'intégration unique d'un gène DfrB en milieu clinique ? et ii) Quels contextes environnementaux et génomiques ont favorisé cette mobilisation entre organismes ?

Il est particulièrement important d'un point de vue de santé publique de s'intéresser aux mécanismes par lesquels les gènes environnementaux peuvent être transférés dans le résistome lorsqu'ils confèrent un quelconque avantage à des bactéries pathogènes.¹⁰ Bien que le développement de nouveaux antibiotiques soit essentiel pour assurer notre capacité à traiter les infections bactériennes, il est tout aussi avisé de réduire au minimum la capacité des bactéries à acquérir des gènes de résistance. Il est donc de première importance de comprendre et répertorier les mécanismes moléculaires et génomiques leur permettant d'étendre l'éventail d'antibiotiques pour lesquels ils ont accès à un mécanisme de résistance. La récapitulation de l'évolution génomique des DfrB au sein d'organismes pathogènes cliniques ira certainement dans ce sens.

Il est pertinent de souligner que la taille des gènes *dfrB*, significativement plus courte que celle des autres gènes présents dans le résistome, représente un atout en biologie synthétique. Li et ses collègues ont créé un plasmide de résistance miniature contenant le gène *dfrB10*, que nous avons rapporté pour la première fois dans le Chapitre 3. Ce vecteur ne mesure que 988 paires de bases, ce qui le rend encore plus petit que le vecteur pUC19 de 2686 paires de bases couramment utilisé. Le gène *dfrB10* a également été intégré au sein du vecteur minimal pUdO pour générer le plus court vecteur de clonage à copie élevée connu, d'une taille de 903 paires de bases.¹¹ Puisque la taille des vecteurs a un impact sur leur efficacité de transformation

dans les cellules,¹² la génération de petits vecteurs représente un avantage significatif dans le domaine de la biologie synthétique.

6.3 L'évolution moléculaire du domaine DfrB

6.3.1 Fonction(s) native(s) du domaine DfrB

Le Chapitre 5, supporté par le Chapitre 4, a permis d'explorer l'évolution naturelle du domaine des DfrB. Ainsi, on a proposé un modèle où l'enzyme homotétramérique des DfrB aurait évolué autour de la capacité à lier une ou plusieurs molécule(s) chargée(s) négativement (Figure 5.3). Dans ce modèle, la dernière étape pour l'émergence d'un domaine DfrB pouvant catalyser la réduction du dihydrofolate est l'évolution de la surface dimère-dimère, qui permet la formation de l'homotétramère caractéristique de la DfrB1.

Pour l'instant, la fonction des homologues de DfrB ne pouvant pas catalyser la réduction du dihydrofolate n'est pas connue. Également, comme proposé au Chapitre 5, et plus particulièrement dans la section 5.3.4, la capacité DfrB et de leurs homologues à réduire le dihydrofolate n'est probablement pas la fonction primaire de ces domaines ; cette dernière constitue peut-être une fonction promiscuitaire. Il est possible que les DfrB servent de modules d'oligomérisation pour la formation de larges complexes avec d'autres domaines, et que ces complexes aient un rôle biologique bénéfique à leur organisme hôte. Également, il est possible que les DfrB lient des molécules à base d'acides nucléiques, voire de l'ADN ou de l'ARN. Par exemple, il est possible d'envisager que le complexe homotétramérique des DfrB se lie à un polymère d'acides nucléiques via une dynamique conformationnelle entre l'homodimère et l'homotétramère. Ainsi, alors qu'on a des indices sur leur fonction native – soit une fonction de liaison aux nucléotides et/ou de multimérisation – ceci reste un pan à explorer. Une première étape serait de tester la liaison de cofacteurs dérivés des nucléotides. On sait déjà que la DfrB1 lie le NAD(P)H, le folate, la dihydrobiopterine et l'ATP-ribose : ceux-ci sont des coenzymes anciens, ou leurs dérivés.^{13,14} Une caractérisation de la liaison du domaine DfrB aux autres coenzymes anciens, tel que le FAD et le coenzyme A, aux nucléotides, et à de courts oligonucléotides permettrait d'avoir une meilleure vue d'ensemble sur cette potentielle fonction de liaison. Les résultats découlant de ces tests de liaison informeraient sur les potentielles voies métaboliques reliées à la fonction native des DfrB et guideraient les prochaines étapes d'investigation.

Ces études d'interaction informeraient également les efforts d'inhibition prenant place au laboratoire Pelletier. Dans le passé, la structure du NADH et du dihydrofolate ont inspiré le design d'inhibiteurs de type bisubstrats pour les DfrB.¹⁵ Ces inhibiteurs sont constitués du groupement adénosyl du NADH et du groupement ptérine du dihydrofolate, favorisant leur liaison. Ils inhibent les DfrB avec une constante d'inhibition dans l'ordre de la dizaine de micromolaire.¹⁵ Si l'on identifie des composés naturels liant les DfrB avec une meilleure affinité, ces derniers pourraient inspirer une nouvelle famille d'inhibiteurs spécifiques. Ainsi, les informations que nous obtiendrons de l'exploration de la fonction native des DfrB

informeront les efforts en chimie médicinale visant à la conception d'inhibiteurs spécifiques et sélectifs aux DfrB impliqués dans le problème de résistance aux antibiotiques.

6.3.2 Ingénierie du domaine DfrB comme modèle d'émergence de la catalyse

Notre étude sur l'évolution du domaine des DfrB use largement d'une approche où les homologues identifiés dans les bases de données sont caractérisés sans modification à leur séquence. Pour comprendre l'importance et le rôle d'éléments clés permettant à ce repliement SH3 de catalyser une réaction chimique (Figure 6.1A), on peut modifier un élément à la fois, et user de la diversité de séquences déjà caractérisées.

D'abord, on peut s'intéresser à la relation entre les sites actifs et leur contexte protéique. Notre exploration de l'espace de séquence a révélé que les sites actifs identiques chez les DfrB sont partagés entre des homologues ayant une haute similarité de séquences (Figure 5.3A) : est-ce que les divers sites actifs pouvant catalyser la réduction du dihydrofolate sont modulables entre domaines DfrB ? Pour répondre à cette question, il serait possible d'interchanger les sites actifs rapportés pour procurer de la résistance contre le triméthoprime. Un homologue dont le site actif est QMIC pourrait être substitué pour le site actif IQNF ou LHVY, et vice versa, sans aucune autre modification du domaine (Figure 6.1B).

Si ces alternatives n'influencent pas la capacité des homologues à catalyser la réaction, on pourrait postuler sur la nature modulaire du tunnel formé par les DfrB. Si ces chimères perdent leur capacité catalytique, on peut alors supposer que les résidus du site actif ont développé des interactions intra- et intermoléculaires. Une analyse plus approfondie de ces interactions chez tous les homologues actifs, jumelée à des tests expérimentaux sur des chimères intégrant les informations relatives à ces interactions serait alors nécessaire pour comprendre ce phénomène. Puisque l'ingénierie au site actif de DfrB1 a révélé au début des années 2000 que quelques combinaisons autres que le motif VQIY peuvent catalyser la réduction du dihydrofolate, il est probable que la première alternative se révèle vraie.¹⁶ Les interfaces monomère-monomère et dimère-dimère pourraient également être interchangées entre homologues, pour explorer leur capacité à être transposés entre domaines (Figure 6.1B).

Pour pousser notre compréhension de ce qui a permis au domaine DfrB de gagner une activité catalytique, on pourrait partir d'un des domaines où l'activité catalytique n'a pas été observée, et qui n'est pas prédit pour créer un tétramère, et intégrer graduellement des éléments de séquences présents chez son homologue actif le plus proche (Figure 6.1C). La caractérisation de la capacité catalytique et de la nature multimérique de ces intermédiaires informerait le modèle évolutif pour l'émergence d'une activité catalytique au sein du domaine DfrB.

Une fois qu'on atteindra une connaissance détaillée des éléments de séquences essentiels aux interfaces protéines-protéines et du site actif, on pourrait entreprendre l'ingénierie de ces éléments au sein de

domaines SH3 ne partageant pas d'homologie de séquences avec les DfrB (Figure 6.1D). Ceci informerait sur le potentiel catalytique des repliements SH3 en général. Si les éléments de séquences essentiels à la catalyse des DfrB sont intégrés à des domaines SH3, est-ce que ceux-ci peuvent catalyser une réaction chimique ?

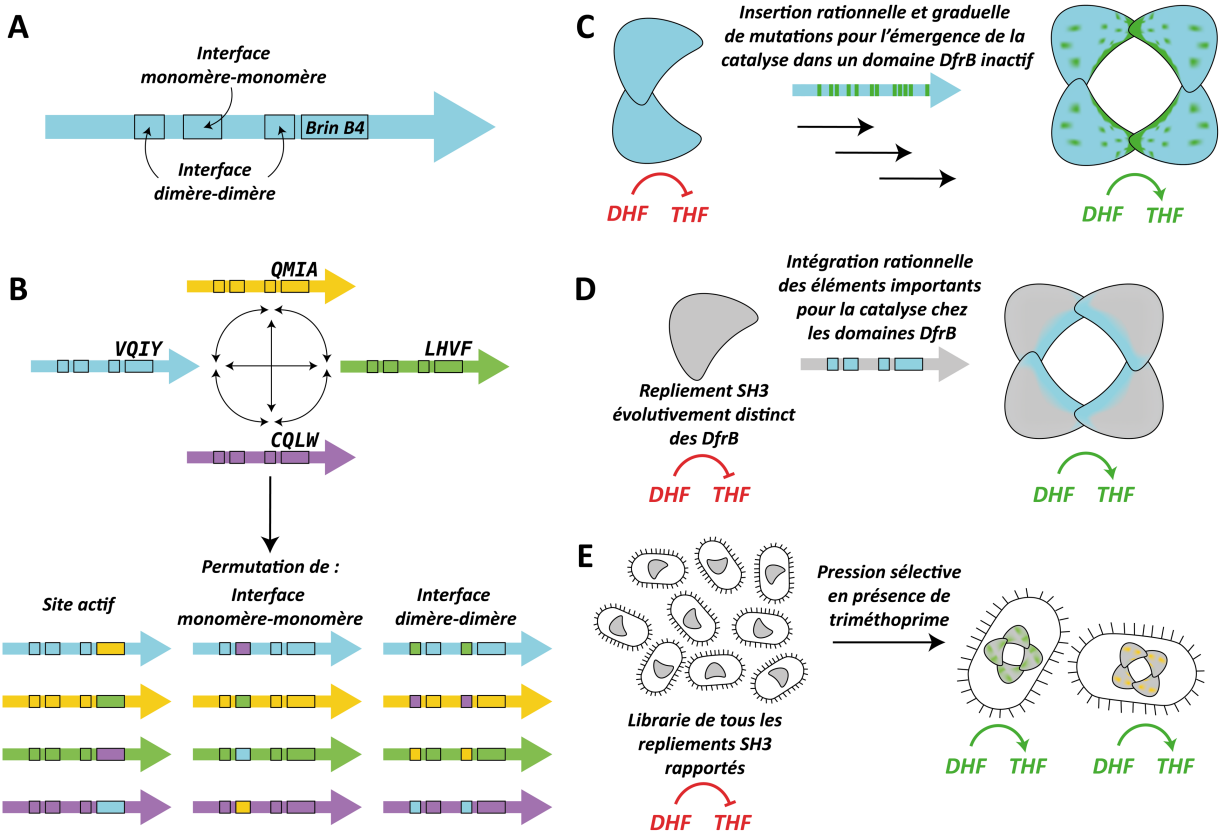


Figure 6.1. Les prochaines étapes pour pousser notre compréhension du domaine DfrB et de sa capacité catalytique.

A. Les éléments principaux pour la formation de l'enzyme DfrB. **B.** Les éléments principaux au sein de quatre homologues de DfrB sont permutés, et leur capacité à catalyser la réduction du dihydrofolate est sondée. **C.** À partir d'un homologue DfrB ne pouvant pas catalyser la réduction du dihydrofolate mais prédit pour former un dimère, des substitutions présentes chez son homologue catalytique le plus proche sont insérées graduellement. L'état de multimérisation et l'activité catalytique des intermédiaires sont caractérisés. **D.** À partir d'un repliement SH3 qui n'est pas un domaine DfrB, les éléments clés chez la DfrB pour former une enzyme fonctionnelle sont insérés jusqu'à ce que ce chimère présente une activité catalytique. **E.** Des bactéries contenant tous les repliements SH3 présents dans les bases de données structurales sont soumises à une pression de sélection par la présence du triméthoprime. Les repliements SH3 identifiés au sein de bactéries résistantes au triméthoprime seront caractérisés enzymatiquement et biophysiquement.

Également, il serait possible, à partir d'une librairie contenant tous les domaines SH3 présents dans les bases de données de structure, de faire une expérience d'évolution dirigée – utilisant le triméthoprime comme pression sélective – pour voir s'il n'y a pas d'autres alternatives pour les domaines SH3 d'évoluer une capacité catalytique (Figure 6.1E). En introduisant des mutations aléatoirement au sein des gènes

codant pour les SH3, et en sélectionnant les itérations pouvant réduire le dihydrofolate, on gagnerait des connaissances sur les propriétés nécessaires à la catalyse dans le domaine SH3, connaissances qui sont présentement restreintes au domaine DfrB.

6.3.3 Pousser l'exploration de l'espace de séquences des DfrB

Le paysage de *fitness* de protéines (pour « *protein fitness landscape* ») est couramment utilisé pour décrire l'évolution d'une fonction chez une protéine. Le *fitness* est généralement défini par l'impact d'une séquence sur le succès reproductif d'un organisme,¹⁷ mais peut être défini autrement dans un contexte de sélection artificielle. Ce cadre conceptuel représente les séquences ayant une aptitude exceptionnelle pour la fonction d'intérêt sous forme de sommets, et celles ayant une aptitude nulle sont représentées sous forme de vallées. Ces paysages peuvent être décrits de multiples façons selon le type de protéine à l'étude, comme par une surface lisse avec un seul pic regroupant les séquences fonctionnelles ou encore par une surface avec de nombreux pics, d'allure accidentée.¹⁸

Les données du Chapitre 5 peuvent être utilisées pour générer ce genre de paysage. Le groupe *I* de l'espace exploré (Figure 5.2; *cluster I*) est particulièrement peuplé. Les 205 séquences le composant partagent un haut niveau d'homologie de séquence entre elles, et représentent des centaines d'autres séquences avec lesquelles elles partagent une identité de séquences encore plus élevée. Le groupe *II*, lui, est beaucoup moins peuplé, avec 70 séquences, alors que la variabilité de séquence s'y trouve plus élevée : certaines séquences partagent un maximum de 42% d'identité de séquences avec le domaine SH3 de l'homologue le plus proche. Ainsi, il serait difficile de générer un paysage de *fitness* pour ce groupe avec les données actuelles. En effet, alors qu'on a caractérisé le potentiel catalytique d'homologues naturellement explorés le composant, le *fitness* des séquences intermédiaires à ces homologues est inconnu. Deux hypothèses peuvent être générées concernant la faible densité de ce groupe. D'abord, il se peut que ce soit dû aux données de séquences auquel nous avons eu accès. La population de séquences peuplant le groupe *II* se retrouvant principalement dans des organismes environnementaux que seule la métagénomique peut révéler, les bases de données auquel nous avons accès ne contiennent qu'une partie de ces séquences. Une autre explication serait que, considérant le fait que la sélection naturelle ne permet pas la fixation de séquences non fonctionnelles, les séquences intermédiaires aux séquences du groupe *II* n'ont pas été identifiées dans les bases de données, car elles n'ont pas été conservées au cours de l'évolution, due au fait qu'elles ne sont pas fonctionnelles.¹⁸

Ainsi, alors que j'ai démontré que des homologues des DfrB évolutivement distants peuvent catalyser la réduction du dihydrofolate avec la même efficacité – et ont donc le même *fitness* en laboratoire (Table 5.1), l'allure du paysage de *fitness* décrivant l'évolution du domaine des DfrB n'est pas connu, puisqu'on ne se sait pas s'il est constitué de deux niveaux plats (le premier composé de séquences non fonctionnelles et

l'autre, de séquences fonctionnelles, Figure 6.2A), ou si les séquences fonctionnelles sont séparées par des vallées de faible *fitness*, où les séquences ne sont pas fonctionnelles (Figure 6.2B).

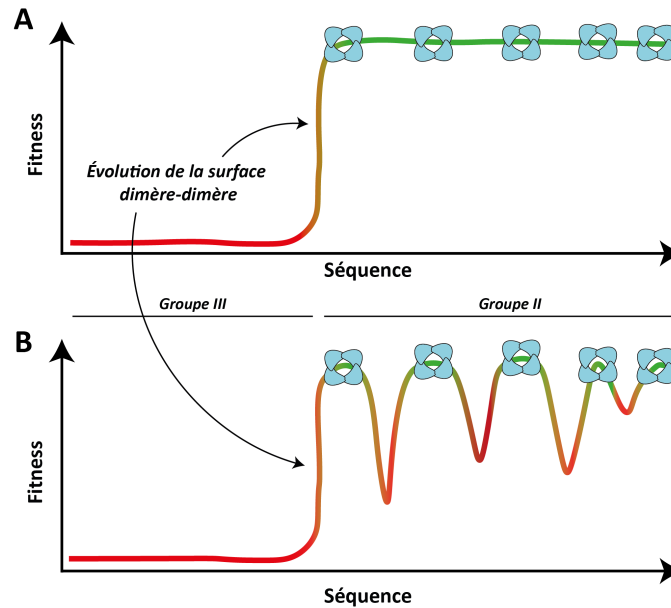


Figure 6.2. Paysage de *fitness* décrivant l'évolution de l'émergence de la catalyse au sein du domaine DfrB.

La notion de *fitness* est associée au potentiel catalytique de chaque séquence étudiée. Les groupes sont identifiés selon la nomenclature définie au Chapitre 5 (*cluster*). Les homologues du groupe II, identifiés et caractérisés dans le Chapitre 5, sont représentés de manière schématique en bleu. A. Les séquences intermédiaires aux séquences des homologues caractérisés démontrent un *fitness* similaire à celui des séquences caractérisées. B. Les séquences intermédiaires aux séquences des homologues caractérisés ne présentent pas de fonctionnalité catalytique.

Pour pallier la faible densité de séquences au sein du groupe II, la méthode permettant la reconstruction de séquences ancestrales (pour « *ancestral sequence reconstruction* ») permettrait de peupler cet espace de séquence. À partir de séquences du groupe II, il serait possible de générer des séquences probables correspondant à leurs ancêtres. Puisqu'on a démontré l'efficacité d'AlphaFold-multimer à prédire la multimérisation d'homologues de DfrB, la prédiction du complexe homotétramérique de ces ancêtres correspondrait à une première étape de criblage. Parmi les itérations générées, certaines seraient synthétisées et testées en laboratoire pour leur capacité catalytique et leur état de multimérisation. Les ancêtres générés pour lesquels AlphaFold-multimer ne prédit pas un tétramère similaire à celui de la DfrB1 seraient favorisés dans cette sélection, puisque ce genre de séquences n'a pas été identifié au sein du groupe II. L'échantillonnage de séquences au sein de ce groupe n'étant pas exhaustif, ceci représente une limitation à l'exactitude des séquences ancestrales qu'on obtiendrait. Pourtant, cette méthode a permis avec succès d'établir l'évolution de l'hétérotétramère de l'hémoglobine à partir de 177 séquences, dont certaines ne partageaient que 36% d'identité de séquences.¹⁹ Ce faisant, on pourrait déterminer lequel des deux scénarios

présentés à la Figure 6.2 représente le mieux le paysage de *fitness* du domaine DfrB pour la catalyse du dihydrofolate en tétrahydrofolate.

Une autre manière de pousser l'exploration et la compréhension de l'évolution du domaine DfrB est par le criblage de certaines propriétés à travers son espace de séquence. Au Chapitre 5, plusieurs propriétés ont été criblées chez ses homologues : la multimérisation, la capacité à réduire le dihydrofolate et la thermostabilité. Ces propriétés ont été sélectionnées, car elles définissaient les DfrB précédemment caractérisées. Pourtant, plusieurs autres propriétés centrales à l'évolution du domaine DfrB peuvent avoir été omises, car elles n'ont pas été identifiées par le passé chez la DfrB1 et ses homologues proches. Ce genre d'étude, où l'on tente de cartographier plusieurs propriétés pour définir leur influence sur l'évolution d'une famille de protéines gagnera par la démocratisation d'analyses permettant de cribler des propriétés à partir de séquences protéiques.

En Annexe 1, on a présenté le rôle central de la dynamique protéique pour l'activité enzymatique, et l'importance d'incorporer celle-ci au sein de pipelines de caractérisation et d'ingénierie d'enzymes. Comme expliqué dans ce chapitre, la génération de données dynamiques – et surtout leur analyse à large échelle – demeure un des freins majeurs à cette intégration. Le cas de la DfrB est un bon exemple : les études expérimentales et computationnelles sur la dynamique du complexe de la DfrB1 révèlent une grande rigidité du squelette peptidique.²⁰⁻²⁵ Ce sont plutôt les substrats, en particulier la queue du DHF, qui démontrent une mobilité importante, à la fois par l'amplitude du mouvement et également pour son rôle essentiel à l'activité catalytique.²⁶ Ainsi, le criblage d'une propriété liée à la dynamique dans l'espace de séquence du domaine de DfrB devrait idéalement inclure ses substrats, puisqu'il n'y a pas de mouvements particuliers au domaine ou au complexe homotétramérique connus. Il est possible que l'évolution du domaine ait mené à une diminution de la dynamique et de mouvements spécifiques, ce qui aurait favorisé la catalyse. À l'heure actuelle, ce genre d'études exploratoires et à grande échelle en dynamique protéique demeure complexe à entreprendre, et consiste en une limitation à notre analyse.

6.4 Questions d'épistémologie

Les résultats de cette thèse soulèvent des questions d'ordre épistémologique, telles que « Qu'est-ce qui définit une DfrB ? » ou « Qu'est-ce qui permet à une protéine de faire partie de la famille des DfrB ? ». Historiquement, les DfrB étaient définies comme des protéines identifiées dans le contexte de la résistance aux antibiotiques, d'une taille de 78 acides aminés, comportant le motif VQIY au site actif, portées sur un plasmide et qui confère une résistance au triméthoprim en réduisant le dihydrofolate en présence de NADPH.²⁷ Cependant, les découvertes du Chapitre 3 et celles découlant des analyses métagénomiques suggèrent que certaines protéines très similaires aux premières DfrB identifiées ne sont pas associées à la

résistance aux antibiotiques et ne sont pas portées sur un plasmide.³⁻⁵ Peut-on les considérer comme faisant partie de la famille des DfrB ?

De plus, dans les chapitres 4 et 5, j'ai caractérisé divers homologues des DfrB qui agissent en tant que dihydrofolate réductases et forment l'homotétramère caractéristique des DfrB. Certains de ces homologues sont de tailles différentes et comportent des domaines protéiques additionnels, tandis que d'autres sont plus courts et ne possèdent pas le terminus amine qui caractérise les DfrB1 à DfrB21. Certains présentent également des variations du site actif, mais parviennent tout de même à catalyser la réduction du dihydrofolate avec une efficacité similaire à celle des premières DfrB identifiées dans les années 1970. En fait, j'ai démontré au Chapitre 5 qu'aucun résidu n'est conservé parmi tous les homologues capables de réduire le dihydrofolate. Ainsi, peut-on considérer ces protéines de tailles et de séquences diverses comme faisant partie de la famille DfrB ?

Dans le cadre de cette thèse, j'ai adopté une terminologie où les DfrB font référence aux DfrB1-DfrB21, qui ont été caractérisées individuellement et rapportées officiellement comme membres de la famille DfrB.³ Toute autre séquence partageant une homologie de séquence avec ces 20 protéines était désignée comme homologue aux DfrB. Plus généralement, je fais référence au « domaine des DfrB » en tant que repliement SH3 dont la séquence est partagée entre homologues identifiés, sur base de profils MMC, ce qui comprend les protéines pouvant catalyser la réduction du dihydrofolate et celles ne catalysant pas cette réaction, et n'étant pas prédites pour former des complexes homomériques similaires à la DfrB1.

À la lumière des connaissances acquises au cours de cette thèse, je propose d'intégrer dans la famille des DfrB – rappelons que cet acronyme vient de « dihydrofolate réductase de type B » – toutes les protéines pouvant catalyser la réduction du dihydrofolate par la formation de l'homotétramère caractéristique décrit au Chapitre 2. Cette définition exclurait toute condition spécifique au niveau de la séquence, de l'origine taxonomique ou de l'échantillon d'où cette séquence a été identifiée. Les DfrB cliniques feraient référence à celles identifiées dans le contexte de l'utilisation d'antibiotiques et/ou des activités humaines ; les DfrB environnementales feraient référence à celles identifiées dans un contexte environnemental non relié aux activités humaines. Les homologues aux DfrB qui ne catalysent pas la réduction du dihydrofolate, comme ceux composant les groupes *III*, *IV* et *V* au Chapitre 5 – seraient alors définies comme des homologues non catalytiques.

6.5 Perspectives

Les travaux présentés dans cette thèse ont clairement démontré la nécessité d'adopter une approche multidisciplinaire pour comprendre, ne serait-ce qu'une infime partie, le monde qui nous entoure. Cela implique de surmonter les frontières souvent érigées entre les domaines traités et enseignés de manière

cloisonnée, et de promouvoir les échanges collaboratifs entre les experts de ces diverses disciplines. Cette étude de l'évolution des DfrB a bénéficié des efforts concertés en génomique, métagénomique, biophysique, enzymologie, ingénierie, biologie structurale et bio-informatique. J'ose espérer que davantage de systèmes protéiques profiteront d'une telle synergie d'expertises, permettant ainsi de révéler la complexité sous-jacente à l'évolution naturelle, une complexité qui peut à son tour inspirer nos efforts en ingénierie.

Tout comme mon doctorat m'a enseigné l'importance d'adopter une approche holistique pour comprendre l'évolution d'une famille d'enzymes, il m'a également appris que cette approche enrichit tout autant le parcours et les connaissances des personnes impliquées.

6.6 Références

- (1) Miłobedzka, A.; Ferreira, C.; Vaz-Moreira, I.; Calderón-Franco, D.; Gorecki, A.; Purkrtova, S.; Jan Bartacek; Dziewit, L.; Singleton, C. M.; Nielsen, P. H.; Weissbrodt, D. G.; Manaia, C. M. Monitoring Antibiotic Resistance Genes in Wastewater Environments: The Challenges of Filling a Gap in the One-Health Cycle. *J. Hazard. Mater.* **2022**, *424*, 127407. <https://doi.org/10.1016/j.jhazmat.2021.127407>.
- (2) Benson, D. A.; Cavanaugh, M.; Clark, K.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Sayers, E. W. GenBank. *Nucleic Acids Res.* **2017**, *45* (D1), D37–D42. <https://doi.org/10.1093/nar/gkw1070>.
- (3) Cellier-Goetghebeur, S.; Lafontaine, K.; Lemay-St-Denis, C.; Tsamo, P.; Bonneau-Burke, A.; Copp, J. N.; Pelletier, J. N. Discovery of Highly Trimethoprim-Resistant DfrB Dihydrofolate Reductases in Diverse Environmental Settings Suggests an Evolutionary Advantage Unrelated to Antibiotic Resistance. *Antibiotics* **2022**, *11* (12), 1768. <https://doi.org/10.3390/antibiotics11121768>.
- (4) Kneis, D.; Berendonk, T. U.; Forslund, S. K.; Hess, S. Antibiotic Resistance Genes in River Biofilms: A Metagenomic Approach toward the Identification of Sources and Candidate Hosts. *Environ. Sci. Technol.* **2022**, *56* (21), 14913–14922. <https://doi.org/10.1021/acs.est.2c00370>.
- (5) Kneis, D.; Lemay-St-Denis, C.; Cellier-Goetghebeur, S.; Elena, A. X.; Berendonk, T. U.; Pelletier, J. N.; Heß, S. Trimethoprim Resistance in Surface and Wastewater Is Mediated by Contrasting Variants of the dfrB Gene. *ISME J.* **2023**. <https://doi.org/10.1038/s41396-023-01460-7>.
- (6) Macedo, G.; van Veelen, H. P. J.; Hernandez-Leal, L.; van der Maas, P.; Heederik, D.; Mevius, D.; Bossers, A.; Schmitt, H. Targeted Metagenomics Reveals Inferior Resilience of Farm Soil Resistome Compared to Soil Microbiome after Manure Application. *Sci. Total Environ.* **2021**, *770*, 145399. <https://doi.org/10.1016/j.scitotenv.2021.145399>.
- (7) Pavlopoulos, G. A.; Baltoumas, F. A.; Liu, S.; Selvitopi, O.; Camargo, A. P.; Nayfach, S.; Azad, A.; Roux, S.; Call, L.; Ivanova, N. N.; et al. Unraveling the Functional Dark Matter through Global Metagenomics. *Nature* **2023**, *622* (7983), 594–602. <https://doi.org/10.1038/s41586-023-06583-7>.
- (8) Berglund, B. Environmental Dissemination of Antibiotic Resistance Genes and Correlation to Anthropogenic Contamination with Antibiotics. *Infect. Ecol. Epidemiol.* **2015**, *5* (1), 28564. <https://doi.org/10.3402/iee.v5.28564>.

- (9) Larsson, D. G. J.; Flach, C.-F. Antibiotic Resistance in the Environment. *Nat. Rev. Microbiol.* **2022**, *20* (5), 257–269. <https://doi.org/10.1038/s41579-021-00649-x>.
- (10) Ebmeyer, S.; Kristiansson, E.; Larsson, D. G. J. A Framework for Identifying the Recent Origins of Mobile Antibiotic Resistance Genes. *Commun. Biol.* **2021**, *4* (1), 8. <https://doi.org/10.1038/s42003-020-01545-5>.
- (11) Staal, J.; Beyaert, R. *Extreme Miniaturization in Plasmid Design: Generation of the 903 Bp Cloning Vector pICot2*; preprint; Synthetic Biology, 2023. <https://doi.org/10.1101/2023.11.29.569326>.
- (12) Hornstein, B. D.; Roman, D.; Arévalo-Soliz, L. M.; Engevik, M. A.; Zechiedrich, L. Effects of Circular DNA Length on Transfection Efficiency by Electroporation into HeLa Cells. *PLOS ONE* **2016**, *11* (12), e0167537. <https://doi.org/10.1371/journal.pone.0167537>.
- (13) Jackson, M.; Chopra, S.; Smiley, R. D.; Maynard, P. O.; Rosowsky, A.; London, R. E.; Levy, L.; Kalman, T. I.; Howell, E. E. Calorimetric Studies of Ligand Binding in R67 Dihydrofolate Reductase. *Biochemistry* **2005**, *44* (37), 12420–12433. <https://doi.org/10.1021/bi050881s>.
- (14) Sanchez Rocha, A. C.; Makarov, M.; Pravda, L.; Novotny, M.; Hlouchova, K. *Coenzyme-Protein Interactions since Early Life*; preprint; Bioinformatics, 2023. <https://doi.org/10.1101/2023.10.28.563965>.
- (15) Toulouse, J. L.; Shi, G.; Lemay-St-Denis, C.; Ebert, M. C. C. J. C.; Deon, D.; Gagnon, M.; Ruediger, E.; Saint-Jacques, K.; Forge, D.; Vanden Eynde, J. J.; Marinier, A.; Ji, X.; Pelletier, J. N. Dual-Target Inhibitors of the Folate Pathway Inhibit Intrinsically Trimethoprim-Resistant DfrB Dihydrofolate Reductases. *ACS Med. Chem. Lett.* **2020**, *11* (11), 2261–2267. <https://doi.org/10.1021/acsmchemlett.0c00393>.
- (16) Schmitzer, A. R.; Lépine, F.; Pelletier, J. N. Combinatorial Exploration of the Catalytic Site of a Drug-Resistant Dihydrofolate Reductase: Creating Alternative Functional Configurations. *Protein Eng. Des. Sel.* **2004**, *17* (11), 809–819. <https://doi.org/10.1093/protein/gzh090>.
- (17) Wright, S. The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution. *Proc VI Int Congr Genet I*, 356–366.
- (18) Romero, P. A.; Arnold, F. H. Exploring Protein Fitness Landscapes by Directed Evolution. *Nat. Rev. Mol. Cell Biol.* **2009**, *10* (12), 866–876. <https://doi.org/10.1038/nrm2805>.
- (19) Pillai, A. S.; Chandler, S. A.; Liu, Y.; Signore, A. V.; Cortez-Romero, C. R.; Benesch, J. L. P.; Laganowsky, A.; Storz, J. F.; Hochberg, G. K. A.; Thornton, J. W. Origin of Complexity in Haemoglobin Evolution. *Nature* **2020**, *581* (7809), 480–485. <https://doi.org/10.1038/s41586-020-2292-y>.
- (20) Pitcher, W. H.; DeRose, E. F.; Mueller, G. A.; Howell, E. E.; London, R. E. NMR Studies of the Interaction of a Type II Dihydrofolate Reductase with Pyridine Nucleotides Reveal Unexpected Phosphatase and Reductase Activity. *Biochemistry* **2003**, *42* (38), 11150–11160. <https://doi.org/10.1021/bi0349874>.
- (21) Krahn, J. M.; Jackson, M. R.; DeRose, E. F.; Howell, E. E.; London, R. E. Crystal Structure of a Type II Dihydrofolate Reductase Catalytic Ternary Complex [†]. *Biochemistry* **2007**, *46* (51), 14878–

14888. <https://doi.org/10.1021/bi701532r>.
- (22) Alonso, H.; Gillies, M. B.; Cummins, P. L.; Bliznyuk, A. A.; Gready, J. E. Multiple Ligand-Binding Modes in Bacterial R67 Dihydrofolate Reductase. *J. Comput. Aided Mol. Des.* **2005**, *19* (3), 165–187. <https://doi.org/10.1007/s10822-005-3693-6>.
- (23) Narayana, N. High-Resolution Structure of a Plasmid-Encoded Dihydrofolate Reductase: Pentagonal Network of Water Molecules in the D_2 -Symmetric Active Site. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62* (7), 695–706. <https://doi.org/10.1107/S09074444906014764>.
- (24) Mhashal, A. R.; Major, D. T. Temperature-Dependent Kinetic Isotope Effects in R67 Dihydrofolate Reductase from Path-Integral Simulations. *J. Phys. Chem. B* **2021**, *125* (5), 1369–1377. <https://doi.org/10.1021/acs.jpcc.0c10318>.
- (25) Stojković, V.; Kohen, A. Enzymatic H Transfers: Quantum Tunneling and Coupled Motion from Kinetic Isotope Effect Studies. *Isr. J. Chem.* **2009**, *49* (2), 163–173. <https://doi.org/10.1560/IJC.49.2.163>.
- (26) Chopra, S.; Lynch, R.; Kim, S.-H.; Jackson, M.; Howell, E. E. Effects of Temperature and Viscosity on R67 Dihydrofolate Reductase Catalysis. *Biochemistry* **2006**, *45* (21), 6596–6605. <https://doi.org/10.1021/bi052504l>.
- (27) Howell, E. E. Searching Sequence Space: Two Different Approaches to Dihydrofolate Reductase Catalysis. *ChemBioChem* **2005**, *6* (4), 590–600. <https://doi.org/10.1002/cbic.200400237>.

Annexe 1. Le prochain défi technologique pour l'évolution et l'ingénierie d'enzymes

Préface

Les protéines sont, par nature, intrinsèquement flexibles. Cette flexibilité a un impact significatif sur la fonction et la capacité d'évolution des enzymes. Afin de mieux comprendre le rôle du dynamisme dans la fonction et l'évolution des enzymes, il est judicieux d'intégrer les informations relatives à leur dynamique moléculaire dans l'ensemble des données concernant leur structure, leur activité, et l'espace de séquence exploré par leurs homologues fonctionnels. Le potentiel de cette intégration a été longtemps limité par la complexité expérimentale et informatique que la caractérisation de la dynamique moléculaire nécessite. Alors que des technologies se développent pour faciliter l'intégration de la dynamique, notre compréhension de la dynamique des enzymes est encore limitée.

Les chapitres précédents ont montré l'impact qu'ont eu les récentes avancées technologiques sur notre capacité à étudier l'évolution d'une famille d'enzymes. Cette annexe se tourne vers l'avant en explorant le potentiel d'intégration de la dynamique moléculaire dans notre compréhension des systèmes enzymatiques, et ainsi dans notre capacité à les évoluer.

La revue de littérature présentée ici a été publiée dans le journal *Protein Engineering, Design and Selection* en 2022, suite à une invitation du journal. Des modifications mineures ont été apportées à la version incluse dans cette thèse. J'ai réalisé la revue de la littérature et participé à la conceptualisation de la revue et à l'écriture de toutes ses sections, conjointement avec Prof. Nicolas Doucet et Prof. Joelle N. Pelletier. J'ai réalisé trois des quatre figures.

Article de revue 2. Integrating dynamics into enzyme engineering

Claudèle Lemay-St-Denis^{1,2,3}, Nicolas Doucet^{1,4,*} et Joelle N. Pelletier^{1,2,3,5,*}

¹ PROTEO, The Québec Network for Research on Protein, Function, Engineering and Applications, Québec, QC, Canada

² CGCC, Center in Green Chemistry and Catalysis, Montréal, QC, Canada

³ Department of Biochemistry and Molecular Medicine, Université de Montréal, Montréal, QC, Canada

⁴ Centre Armand-Frappier Santé Biotechnologie, Institut National de la Recherche Scientifique (INRS), Université du Québec, Laval, QC, Canada

⁵ Chemistry Department, Université de Montréal, Montréal, QC, Canada

*Correspondence: joelle.pelletier@umontreal.ca ; nicolas.doucet@inrs.ca

Protein Engineering, Design and Selection

DOI : [10.1093/protein/gzac015](https://doi.org/10.1093/protein/gzac015)

© Oxford University Press 2022

A1.1 Abstract

Enzyme engineering has become a widely adopted practice in research labs and industry. In parallel, the past decades have seen tremendous strides in characterizing the dynamics of proteins, using a growing array of methodologies. Importantly, links have been established between the dynamics of proteins and their function. Characterizing the dynamics of an enzyme prior to, and following, its engineering is beginning to inform on the potential of ‘dynamic engineering’, i.e., the rational modification of protein dynamics to alter enzyme function. Here we examine the state of knowledge at the intersection of enzyme engineering and protein dynamics, describe current challenges and highlight pioneering work in the nascent area of dynamic engineering.

A1.2 Why protein dynamics are relevant to engineering catalytic function

Enzyme engineering has become a well-established and broadly practiced approach to modify naturally evolved enzymes for a wide range of applications. Enzymes have been engineered to display improved stability at various temperatures, pH and solvents ¹, to react with non-native substrates or cofactors ² and to catalyze reactions never observed in natural enzymes ^{3,4}, both as stand-alone catalysts ^{5,6} or as part of complex systems ⁷.

Enzyme catalysis requires a multitude of motions of widely differing rates and amplitudes that can involve the entire body of the protein. The catalytic cycle requires the approach of a substrate in a productive conformation into an active-site environment that explores conformational space to become appropriately poised to undertake a catalytic transformation event with speed and precision, followed by release of products and cofactors, and resetting the enzyme to begin a new cycle ⁸. Myriad motions of side chains, backbone or domains can be solicited throughout this process ⁹. It follows that engineering motions in an enzyme has the potential to productively alter activity.

Here, we consider how knowledge of enzyme dynamics provides the possibility to bring enzyme engineering to a higher level. How will observing and understanding the impact of protein motions on catalysis translate into more powerful approaches to modify enzymes for our needs? How can our understanding of motions translate into methods to engineer better enzymes? We focus on the sector of enzyme engineering where new or improved catalytic features are developed in a preexisting enzyme. In terms of complexity, this lies between the better-understood area of enzyme stabilization and pioneering efforts in *de novo* creation of enzyme activity ^{6,10,11}.

Enzyme engineering generally starts with an enzyme of known sequence; its structure has been resolved or modeled and activity assays relying on one or more substrates or analogs are available. Sequence diversity is introduced into the starting enzyme template ¹² followed by screening for the desired new property. If

screening is successful, the desired new features may be identified and optimized in a few rounds of sequence diversification and screening (Figure A1.1A).

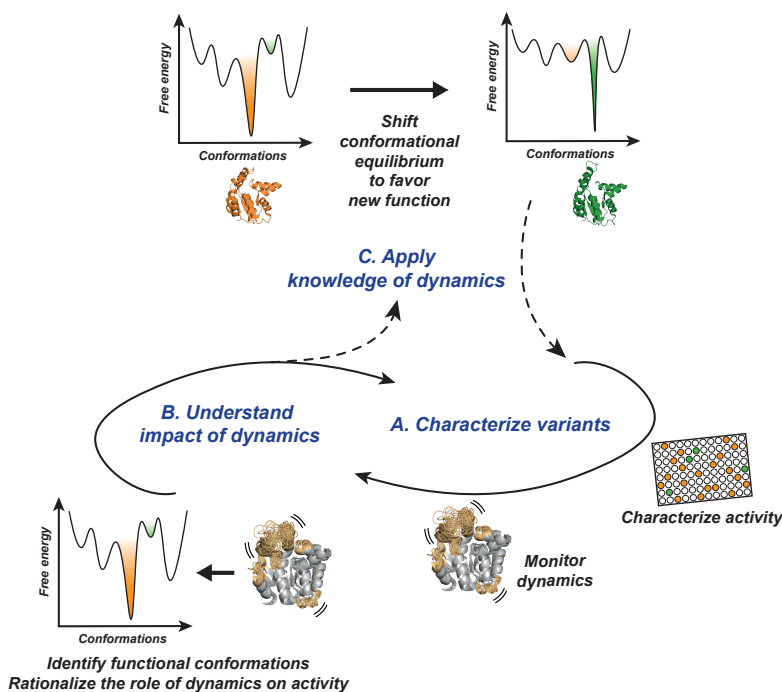


Figure A1.1. State-of-the-art and current challenges in dynamic engineering.

A. Enzyme engineering minimally entails sequence diversification followed by assessment of catalytic properties. Additional assessment of physical properties may include characterizing dynamics. **B.** Taking it one step further, comparing the dynamic properties of variants or homologs exhibiting different phenotypes can allow rationalizing the role of dynamics in enzyme function. **C.** Dynamic engineering entails applying knowledge gained in (B) to guide the next round of sequence modification (C → A → B, etc.).

Rounds of enzyme engineering require determination of catalytic properties under relevant conditions and generally include structural or biophysical characterization of the engineered variants. Structural determination is by far the most widely used approach to establish the link between modifications made to an enzyme and new function or properties. Whereas structure can inform on many important elements of function, temporal fluctuations within secondary and tertiary structure are essential to allow conformational search of the transition state upon ligand binding, for successful chemical transformation or to promote product release.

These conformational changes are intrinsic to enzymes. They are crucial for activity¹³ and can even constitute the rate-limiting step of the catalytic reaction^{14,15}. From the accommodation of the substrate in the active site to allosteric regulation and product release, these movements participate in the catalytic turnover^{16,17}. Fast femto-second (fs) bond vibrations and pico- to nano-second (ps-ns) side-chain motions

are complemented by slower large-scale motions of nanometer scale that occur over micro- to milli-seconds (μs - ms). Faster timescale dynamics are involved in chemical bond breaking and formation whereas slower residue motions and structural rearrangements are involved in events such as ligand (re)positioning, binding, and/or release that limit the rates of catalytic turnovers^{8,9}. Since most enzymes exhibit rate constants ranging between 1 to 10,000 s^{-1} , motions occurring on the μs - ms timescale are of particular interest for modulating rate-limiting steps in enzyme catalysis^{18,19}. Altogether, these motions coexist on multiple timescales, generating a vast number of potential conformations that are populated according to their relative free energy, together describing the energy landscape of an enzyme. Populating the energy landscape of an enzyme and of its engineered variants is increasingly feasible and informs on the effects of sequence modification on dynamics (Figure A1.1B)²⁰.

A1.2.1 The conformational landscape defines enzyme function

Highly specialized enzymes are described by an energy landscape where the conformational minimum responsible for the main reaction pathway is well-defined and highly populated, its free energy being significantly more favorable than other minima. In contrast, promiscuous enzymes exhibit several conformational minima separated by low energy barriers. These alternative minor states can be functionally active and give rise to promiscuous reactions upon conformational sampling, as they allow exploration of the perfect or near-perfect active site geometry for a novel reactivity^{21,22}. Engineering a function can take advantage of the evolvability allowed by conformational flexibility, as well as minimizing nonproductive dynamics for improved catalytic efficiency^{17,23,24}.

Importantly, the introduction of even a single amino acid substitution can induce significant population shifts in the conformational landscape. Their impact on function ranges broadly, to the extent of promoting a new activity^{21,22,25}. This is consistent with protein dynamics constituting an essential aspect of natural enzyme adaptation and emergence of new catalytic functions²⁶ and to adaptation of enzymes to a changing environment^{14,27-30}. Evolution of new functions has been observed in conjunction with an enrichment of pre-existing conformational and catalytically relevant sub-states^{13,31,32}. For instance, alteration of the conformational landscape to favor a stable active-site configuration and minimize ‘unnecessary dynamics’ was observed in variants on the evolutionary trajectory from a phosphotriesterase to an arylesterase³¹ and, similarly, in the evolved variant of a computationally designed Kemp eliminase³³.

It is essential to point out that not all motions play an observable role in catalysis. Indeed, in a set of engineered, chimeric β -lactamases, fast (ps-ns) and intermediate (ns- μs) dynamics were mostly conserved and may be essential for function. In contrast, slow motions (μs - ms) differed widely between the laboratory-engineered variants yet were functionally tolerated³⁴. The engineered chimeras displayed kinetics similar to the native proteins they originated from, indicating that the rate-limiting steps had not been significantly

altered. In another example, motions in the ubiquitous dihydrofolate reductases promote catalysis but have evolved in distinct patterns. In *E. coli* dihydrofolate reductase, the 15-residue Met20 loop transitions between the closed Michaelis complex and a conformation that occludes the active site. This large motion promotes cofactor and product dissociation and is rate-determining. In contrast, the homologous Met20 loop of human dihydrofolate reductase does not exhibit this slow loop motion, its rate being otherwise limited. Nonetheless, their hydride transfer step appears to rely on conserved fast motions³⁵⁻³⁷.

This immediately highlights an important limitation in current approaches to enzyme engineering: despite enzyme function being modulated by dynamics, the acquisition and, in particular, the application of dynamic knowledge are not effectively integrated into enzyme engineering workflows (Figure A1.1A-C). Over the past decade, characterization of protein dynamics has gained in importance. A broad array of experimental and computational approaches focusing on local effects – generally in the active-site area – or on global, long-range dynamics have become established. Characterization of protein dynamics has primarily relied on an integrative approach to structural biology^{38,39}, combining methods that sample multiple timescales and provide information on both averaged ensembles and atomic-scale details of specific conformers. X-ray crystallography and NMR spectroscopy are typically used in complementary fashion to offer atomic-scale structural resolution (X-ray) and a measure of how proteins shift from one conformation to another in solution (NMR). Room temperature and time-resolved X-ray crystallography have also broadened the horizon of functional dynamic information extracted from X-ray diffraction⁴⁰⁻⁴³. In parallel, NMR methods such as spin-relaxation, residual dipolar couplings, $R_{1\rho}$, Carr-Purcell-Meiboom-Gill (CPMG) relaxation dispersion, chemical exchange saturation transfer (CEST) and hydrogen/deuterium (H/D) exchange have allowed atomic-scale characterization of an unprecedented breadth of motions and their structural/temporal localization in proteins⁹. Among others, these experimental methodologies have been complemented by newer developments in sub-second time-resolved mass spectroscopy, cryo-electron microscopy (cryo-EM), hydrogen-deuterium exchange mass spectrometry (HDX-MS) and single-molecule fluorescence resonance energy transfer (FRET)⁴⁴⁻⁴⁹.

Computational methodologies that simulate protein motions on various timescales relevant to catalysis are emerging, diversifying the dynamic descriptors and increasing the quality of the conformational landscape describing a given protein system. Invisible conformational sub-states not yet amenable to experimental scrutiny can also easily be observed using MD methods^{39,50,51}. Importantly, online tools make performing MD simulations highly accessible to the non-expert⁵²⁻⁵⁴.

Yet even as dynamics are increasingly characterized and their impact on catalysis better understood, how can this knowledge feed back into engineering protocols to streamline the workflow and increase the likelihood of achieving significant improvement of catalytic features? That ultimate step in the complete

cycle of dynamic engineering, defined as the rational modification of protein dynamics to alter enzyme function, remains the most challenging (Figure A1.1).

A1.3 How knowledge of protein dynamics could enrich enzyme engineering strategies

A1.3.1 Dynamics as a design tool

As a design tool, uncovering dynamic sites and allosteric pathways linking spatially distant sites within an enzyme structure could identify hidden residue nodes and allosteric sites that act as ‘dynamic switches’ to positively or negatively modulate enzyme function⁵⁵. Understanding the overall ‘dynamic fingerprint’ of distinctive protein folds thus offers emerging targets to specifically modulate and/or design new enzyme activities. Indeed, the ensemble view chosen to represent the body of dynamics in a system, as in Figure A1.1C, can be otherwise expressed as an allosteric network. One view (Figure A1.1C) represents an ensemble of conformers in solution while the other (Figure A1.2) maps the dynamic network on a single structure. Exploiting these dynamic nodes, grafting flexible domains to rigid scaffolds, modulating the dynamics of tunnel opening or shifting equilibrium between conformational populations as a means to modulate function will become the new focus of protein engineering^{8,56-58}. The body of knowledge presented here reveals the potential of modulating dynamics as a means to engineer enzyme activity and informs us that dynamics should be a key target of enzyme engineering.

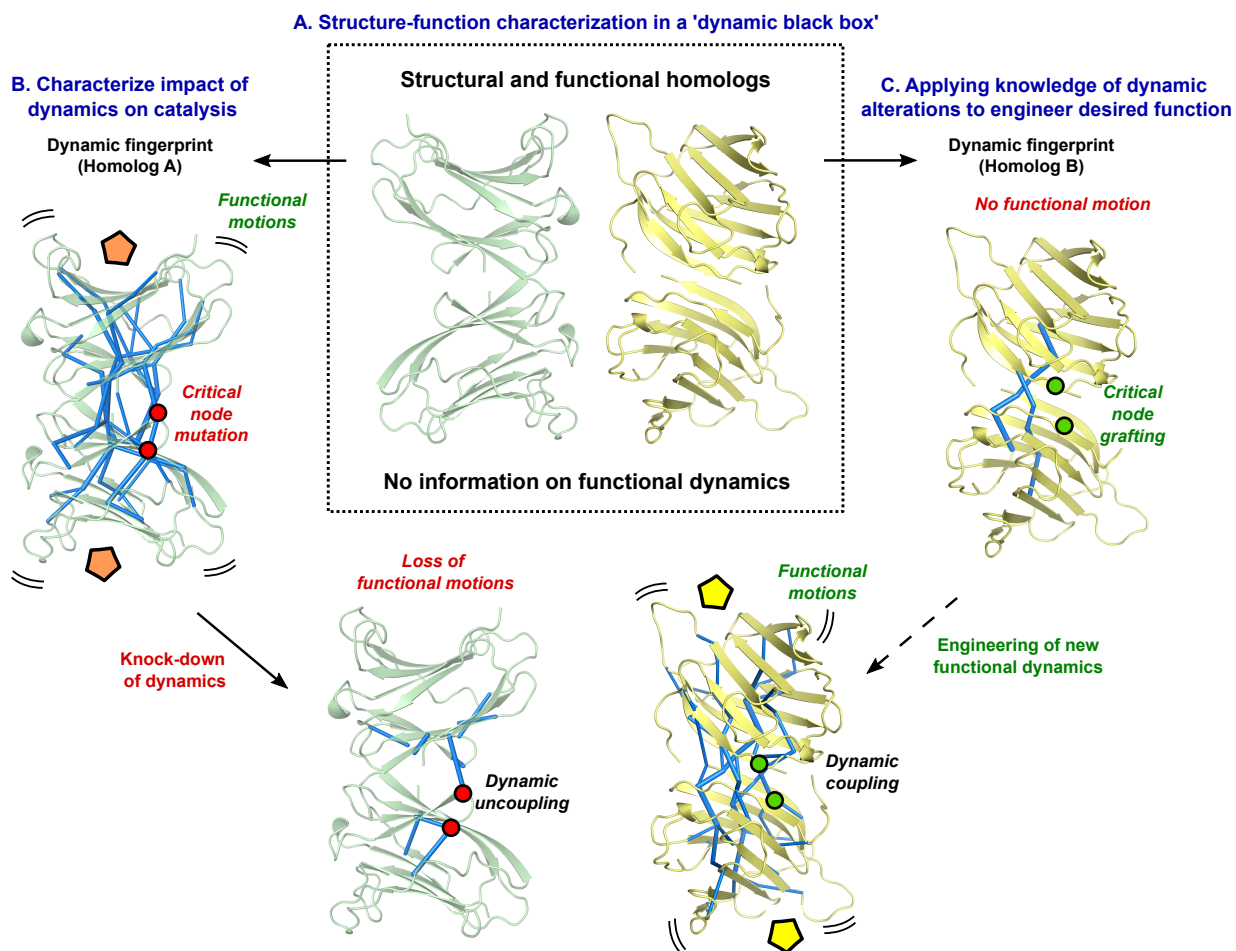


Figure A1.2. Integrating dynamic engineering into enzyme design.

A. The ‘dynamic black box’ of traditional biochemical characterization. Structure alone rarely suffices to understand why functional homologs exhibit distinct catalytic properties or efficiencies. **B.** Dynamic characterization uncovers hidden allosteric behavior or conformational exchange experienced by enzymes. Microstates sampled in solution form discontinuous dynamic sectors within the 3D structure and interconnect to broader independent networks that form global communication pathways (blue wires). Communication between independent dynamic sectors is further controlled by critical nodes of cross-talking residues (red dots) that modulate long-range functional motions that have evolved to maintain efficient catalysis (orange substrates)²⁶. Knocking down a critical node (red dots) through various perturbations (mutation, allosteric modulation, etc.) uncouples native dynamics and perturbs functional motions, resulting in reduction or loss of activity (full arrow)¹²⁰. **C.** Despite high structural similarity, dynamic behavior is unique and transcends sequence homology between functional homologs. Not all structural homologs natively rely on functional motions to promote efficient catalysis. Identifying critical nodes for *de novo* design of dynamics (green dots, dashed arrow) would allow ‘allosteric programming’ of functional motions (blue wires) to improve existing function or create new catalytic activities (yellow substrates).

While all proteins move along the space-time continuum, they do not all exploit catalytic or functional motions for the same purposes (Figure A1.2). For instance, a flexible active-site loop in one dihydrofolate reductase appears to have evolved to promote product release, whereas rigidity within the same active-site

loop in a structural homolog is required to maintain proper substrate positioning during catalysis; this also holds true for some RNase A homologs ^{26,27}. In both enzyme families, loop dynamics are essential for optimal catalysis, but spatial directionality of motions, time frames, and length scales are distinct among the different members of the family.

Emerging technologies to interrogate and characterize dynamic events occurring on multiple timescales within distinct protein scaffolds continue to highlight the importance of protein motions in enzyme function. Chen and Schwartz recently investigated fast timescale dynamics in a family of laboratory-evolved Kemp eliminases ⁵⁹. A *de novo* designed Kemp eliminase had previously served as a starting point for directed evolution, resulting in a 2000-fold improvement of catalytic efficiency ⁶⁰. The authors characterized femtosecond motions termed ‘rate promoting vibrations’ in variants selected throughout that laboratory proxy for an evolutionary path. They observed increasingly dynamic active sites, reflecting the capacity of the new motions to promote the rate of the chemical transformation in the catalytic process ⁶¹.

In that example, Chen and Schwarz suggest that the lack of dynamic design is a limitation in enzyme engineering ⁵⁹. Indeed, the catalytic cycle presents numerous opportunities for dynamic engineering. Substrate and cofactor binding allow chemical transformation, followed by product release. Each of these steps is promoted by conformational changes that could potentially be favored through engineering of relevant positions or regions ⁶². However, accumulating experimental evidence or demonstrating the functional importance of selected dynamic events does not automatically translate into successful design predictions. Indeed, modulating dynamics to our advantage is a much more daunting task than simply observing or characterizing it (Figure A1.1, Figure A1.2).

Engineering greater enzyme stability has been addressed using a variety of approaches, often based on sequence alignment ^{63,64} as well as structural information ^{65–68}. Stability is thus shown to be successfully addressed on the basis of sequence and structure data, allowing the implementation of machine learning to accelerate enzyme stabilization ⁶⁹. Nonetheless, protein dynamics underlie stability and have been specifically targeted in some computational efforts ⁷⁰. Upon identification of flexible regions in a highly thermostable carbonic anhydrase using MD simulations, Parra-Cruz *et al.* designed mutations aimed to increase compactness ⁷¹. From the most promising variants evaluated by FoldX ⁷², one variant was predicted to be stabilizing, significantly reducing flexibility. The free-energy landscape of the system was significantly altered yet root-mean-square fluctuation (RMSF) of active-site residues remained similar, suggesting that conservation of active-site flexibility is essential for maintaining native-like enzyme activity. That report illustrates how knowledge of the dynamics was rationally integrated into a workflow to engineer a physical property of the enzyme.

Prediction of stability is now readily accessible via a growing number of online servers offering varying degrees of accuracy⁷³. While it may be tempting to assume that dynamic events act as accurate predictors of protein stability, the relationship between protein flexibility and stability, both described by short and long-range interactions, has been shown to be complex^{74,75}. Among other factors, atomic-scale flexibility is temporally and spatially relative, such that flexibility on one timescale can coexist with rigidity on another, and differ between homologs displaying similar activity^{34,75}. Better coverage of protein dynamics in a diversity of systems will contribute to a better understanding of the role of residue interactions and motions on several timescales in relation with stability⁷⁶.

Characterization of dynamics has also allowed identification of a residue involved in substrate binding in an area distant from the catalytic site. Ebert and coworkers generated a free energy map describing the trajectory of a substrate traveling through a tunnel into the active site⁷⁷. The authors identified a specific residue within the tunnel that formed important hydrogen bonding interactions with the substrate. Experimental validation of its role by mutagenesis resulted in knocking down catalytic efficiency (k_{cat}/K_M) 8-fold. Thus, information was gained but catalytic function was not improved or diversified by the use of dynamic information (Figure A1.1B).

A1.4 The state-of-the-art in dynamic engineering of enzyme function

To this day, no unique predictive approach has successfully uncovered universal rules governing how dynamics can be reliably exploited to improve enzyme engineering workflows. Below, we present key examples that illustrate successes and highlight challenges in understanding the role of dynamics in enzyme activity, ultimately implementing that knowledge to engineer enzymes.

In the current state-of-the-art, only a few examples have garnered sufficient understanding of the relationship between enzyme dynamics and function to allow successful rational improvement of catalytic activity by implementing knowledge-based dynamic design⁷⁸. Schenkmyerova and colleagues presented a case study where rational modulation of dynamics in a targeted region of the ancestral luciferase AncLuc resulted in 7,000-fold improvement in catalytic activity⁷⁹. As described in Figure A1.3, their workflow began by acquiring functional knowledge of the system by screening for luciferase activity in a library of random insertions and deletions of a reconstructed and catalytically versatile ancestral luciferase, AncLuc. They identified a variant exhibiting a 124-fold increase in catalytic efficiency of the poorly active ancestor to yield variant AncLuc^{Ins1}. By characterizing stability and luciferase/haloalkane dehalogenase activity, the authors highlighted important activity predictors and differences in conformational flexibility and substrate binding. Complementary hydrogen-deuterium exchange mass spectrometry (HDX-MS, s to min dynamics) and molecular dynamics (MD) simulations (μ s dynamics) were then undertaken to identify differences in dynamic behavior that explain differences in activity between AncLuc, AncLuc^{Ins1}, and the modern and

highly active ModLuc. Comparative analysis suggested a correlation between luciferase activity and active-site dynamics in the $\alpha 4$ helix and in loop L14. By grafting the flexible L9- $\alpha 4$ fragment of ModLuc into AncLuc, the dynamics of the chimera's $\alpha 4$ region significantly increased and led to improvement of catalytic efficiency by more than three orders of magnitude relative to AncLuc. This is a ground-breaking example of how dynamic information can be applied to enzyme engineering and design.

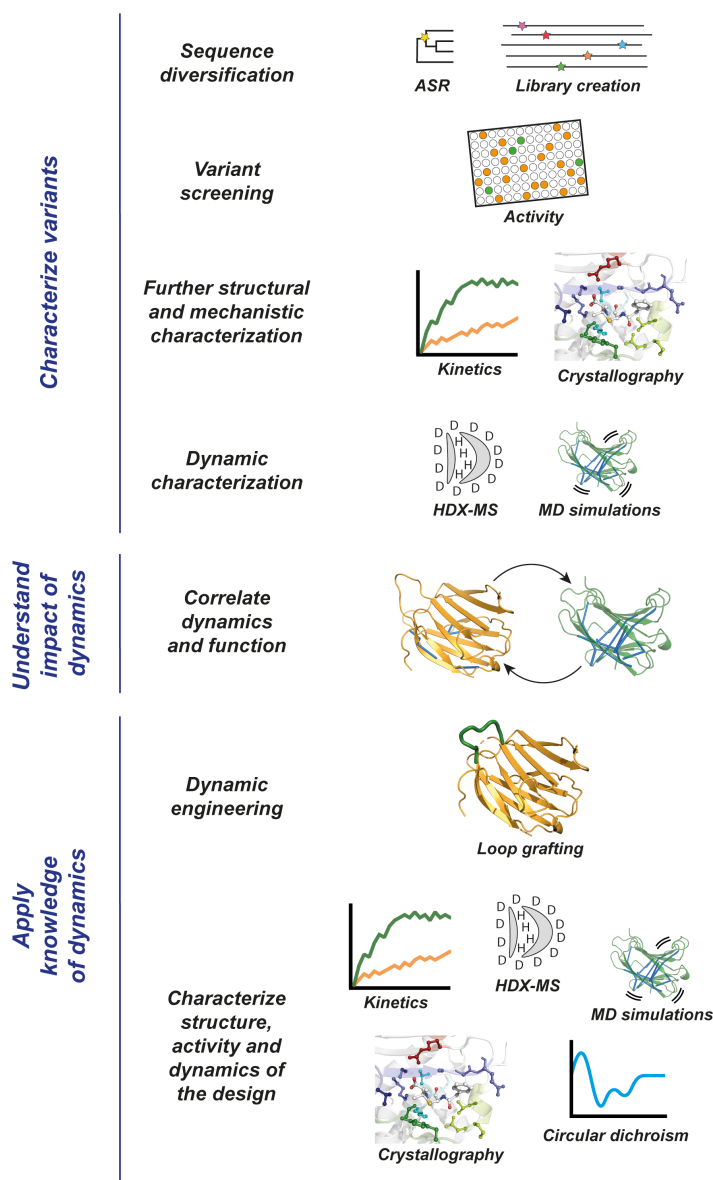


Figure A1.3. Proposed workflow for dynamic engineering based on reference 79.

Using a promiscuous enzyme as starting point, a library of variants is produced. Following screening for the activity of interest, the most promising variants are characterized for their kinetic, structural and dynamic properties. Dynamic information is then linked to catalytic function in the variants, to rationalize the relationship between sequence modification, dynamics and enzyme activity. Finally, this knowledge is implemented in engineering of enzyme function by modulating dynamics. Characterization of the engineered variants will inform the next round of engineering. Figure inspired by ⁷⁹.

Naturally evolved homologs can also provide valuable knowledge on the relationship relating enzyme dynamics to function through identification of conformational similarities and differences between homologs. Using a workflow similar to that shown in Figure A1.3, Bata et al. recently investigated the dynamic behavior of modern 5-methylene-3,5-dihydro-4H-imidazol-4-one (MIO)-family enzymes to engineer the substrate tunnel⁸⁰. The team investigated ligand egress to gain insight into the product release mechanism by performing random acceleration molecular dynamics (RAMD) in structures of eukaryotic and prokaryotic phenylalanine aminomutase (PAM) and phenylalanine aminolyase (PAL) enzymes. Two main pathways for product release were proposed, which turned out to be fully or partially conserved in other MIO-enzymes with the notable exception of the (*S*)-selective PAM of *Pantoea agglomerans*. The two residues responsible for the absence of tunnels in this (*S*)-selective PAM were integrated into a homologous eukaryotic PAM, successfully altering activity and selectivity to yield a previously unreported (*S*)- β -aminolyase activity.

Ancestral sequence reconstruction (ASR) is an increasingly popular tool to explore the sequence space sampled by nature and to investigate how dynamics could have been tailored to enzyme function throughout evolution⁸¹. An important advantage of characterizing the ancestors and modern homologues of a specific enzyme is that it allows exploration of dynamic impacts on the complete protein sequence, as opposed to rationally focusing on the active site environment²⁴. This is necessary to identify allosteric sites that mediate long-range dynamic events that modulate activity. Previous work on ancestral tryptophan synthase brought to light the stand-alone activity of the last bacterial common ancestor (LBCA) of TrpB, as opposed to the allosterically-dependent modern TrpB⁸². Using this LBCA TrpB, Maria-Solano and colleagues used a shortest path map (SPM) protocol to identify 68 “possible hotspots that potentially regulate the enzyme conformational dynamics”, and thus enzyme activity⁸³. Sequence comparison between LBCA TrpB and the oldest allosteric-dependent ancestor (ANC3 TrpB) extracted by ASR yielded 42 potential hotspots. Through sequence comparison and common SPM hotspots, they identified six residues playing probable roles in the stand-alone function. Only one of these residues was in the active-site cavity; nonetheless, designing these residues into ANC3 TrpB generated a 7-fold increase in k_{cat} , thus improving the stand-alone function by means of dynamic engineering.

Fast protein motions essential to catalysis can also be engineered. Building on the knowledge of the catalytic mechanism and transition state of the purine nucleoside phosphorylase (PNP), Zoi and colleagues identified mutations that could alter transition state formation by means of isotopic diversification⁸⁴. Although the catalytic cycle was not accelerated following their engineering cycle, the chemical transformation step was changed by the rational introduction of mutations that modulate fast motions. These examples all highlight

how deep dynamic characterization can pinpoint dynamics that modulate catalytic events, and the feasibility of engineering dynamics to modify catalysis.

A1.5 Outlook: How can dynamic engineering become widely implemented?

A brief history of method development for enzyme engineering is sketched out in Figure A1.4. We have seen a rapid increase in our understanding of the interrelation between enzyme structure and function as a result of methodological developments in creation and screening of more complex libraries. The introduction of dynamics in this process is nascent: the rulebook of dynamic engineering has yet to be written. The amplitude of increase in enzymatic activity that may result from the implementation of protein dynamics into the enzyme engineering workflow – in combination with other engineering methods, in particular to accelerate computational enzyme engineering – is currently uncharted.

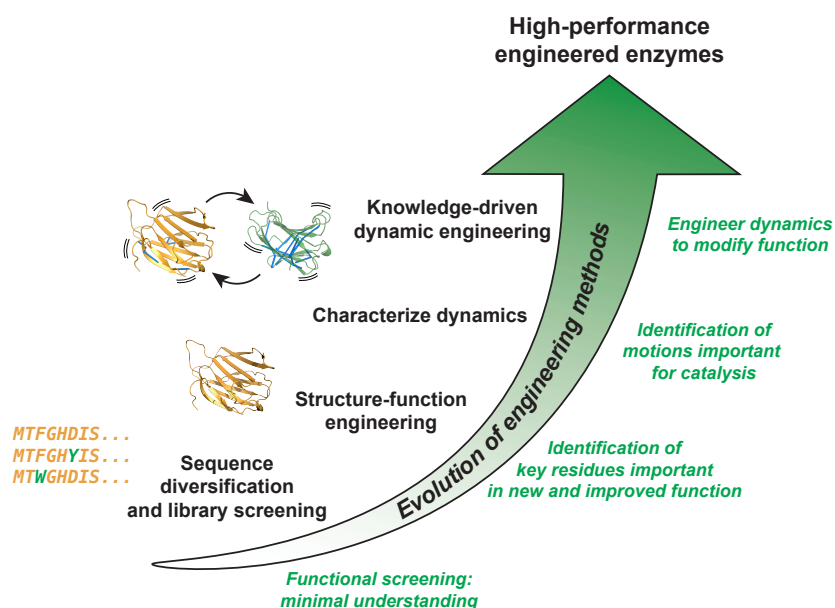


Figure A1.4. A brief history of enzyme engineering methods.

Over the past decades, the actions undertaken to engineer enzymes (black text) have evolved hand-in-hand with progress in our understanding of enzyme structure, function and dynamics, and their interconnection (green text). Computational advances and machine learning as well as development of dynamic databases will facilitate dynamic engineering of enzymes in the next wave of advances.

Even as library design strategies for directed enzyme evolution are increasingly facilitated by the development of molecular biology tools and commercial services¹², ingenious strategies for high-throughput screening are continually emerging to address diverse challenges. Cell-free translation can overcome expression issues⁸⁵ and *in vitro* microcompartments can be compatible with microfluidics to boost the scale of screening by several orders of magnitude^{86,87}, particularly when coupled with powerful

analytical tools such as cell-sorting or ultra-rapid mass spectrometry^{88,89}. Biosensor-based approaches for in-cell selection are also increasingly being developed, broadening the scope of enzymes that can be engineered in a complex cellular environment^{90,91}. In addition, widespread genome sequencing and the advent of metagenomic approaches increasingly provide unprecedented access to natural genetic diversity, broadening our knowledge of catalytically-competent sequences⁹²⁻⁹⁴.

In contrast, dynamic engineering of enzyme function is in its early days. Acquiring knowledge of enzyme dynamics will add depth and breadth to our capacity to understand catalysis and to engineer enzymes. Below, we discuss remaining hurdles to the implementation of dynamic information into decision-making, highlighting promising approaches and technologies that stand to make an impact in this rapidly developing area. We also propose avenues to accelerate the wider implementation of dynamic engineering of enzymes.

A key hurdle to effective implementation of dynamic engineering is the lack of universality in the relationship between enzyme catalysis and molecular scale dynamic events, thus imposing system-specific characterization. Currently, experimental and computational characterization of enzyme dynamics is highly resource demanding, such that comprehensive characterization of dynamics has been described for a strikingly limited number of systems^{34,95,96}. It follows that the collective knowledge base of enzyme dynamics needs to be massively increased (Figure A1.1A) for the impact of dynamics on enzyme function to be understood for specific systems (Figure A1.1B), enabling dynamic engineering (Figure A1.1C). With more knowledge, trends should become apparent, and a diversity of methods for dynamic engineering of enzymes should emerge.

To make dynamic characterization more broadly accessible and more informative, it will be essential to enable higher throughput dynamic studies of new systems and of homologs within systems; to achieve extensive all-atom coverage; and to facilitate the description of longer timescales (larger motions). The aforementioned advances in dynamic resolution by crystallography, NMR, mass spectrometry, electron microscopy and FRET are becoming firmly established and faster to undertake as instrumentation, reagents and automated data-processing software become more broadly accessible. Recent progress has been achieved in several experimental fields to enable atomic-level dynamic characterization of a broader array of protein systems. For instance, selective isotopic labeling now provides high-resolution NMR dynamics analysis of mega-Dalton complexes⁹⁷, complemented by advances in single-molecule cryo-EM dynamic investigation⁴⁹. Computational tools for protein engineering are increasingly accessible to non-specialists via online servers. Some have integrated protein dynamics in their predictive workflow, often in the form of normal mode analysis⁹⁸⁻¹⁰⁰. Access to computational resources remains limiting (who wouldn't want more computational capacity?), hindering the establishment of long MD calculations of large numbers of

enzyme variants in a high-throughput manner; progression of computational resources will necessarily improve even as calculations will increase in complexity.

A further challenge is that mapping the unique dynamic fingerprint of individual biocatalyst systems currently provides no predictive path to successful dynamic engineering (Figure A1.1C). Principal component analysis (PCA) is often used to reduce the number of dimensions of a trajectory, extracting the most relevant elements to describe protein dynamics¹⁰¹. User-friendly tools to facilitate the analysis of MD trajectories using PCA will be helpful⁹⁸. Nonetheless, analysis of complex dynamic datasets and algorithms to link the observation of dynamic events to catalytic events – whether acquired by experimental or computational methods – require more development.

Machine learning (ML) approaches, trained to detect emerging patterns in annotated datasets and then predict properties in new datasets, emerge as promising tools to apply dynamic knowledge toward engineering enzyme function^{102–104}. A recent review by Mazurenko, Prokop and Damborski provides an excellent overview of the state-of-the-art in implementation of ML algorithms for enzyme engineering yet, tellingly, only mentions the potential for dynamics to be considered in the future¹⁰⁵. In an example of ML that includes dynamics, Barros and colleagues successfully differentiated high-affinity binders from non-binders in an unsupervised ML approach fed features acquired during MD simulations¹⁰⁶. They identified defined MD descriptors that most effectively distinguished binding from non-binding designs using PCA and clustered the designs using the k-means algorithm. Upon generating three distinct clusters, only one contained all binding designs. This demonstrates how automation can substantially accelerate the analysis of the extensive dynamic information to extract the core elements coupling dynamic alteration to modified function, yet it did not directly address catalysis.

Once the enzyme engineering field will have validated its capacity to successfully integrate protein dynamics into design of function, the next goal should be to incorporate dynamics in *de novo* enzyme design¹⁰⁷. While active site pre-organization is certainly of the utmost importance, consideration of correlated internal motions is proving to be critical for the design of efficient catalysts^{41,108}. Integration of dynamics into *de novo* enzyme design is, at this point, largely unexplored^{41,109,110}, but will eventually demonstrate a real understanding of how enzymes – inherently dynamic molecules of great complexity – function and evolve¹¹.

A further barrier to broad implementation of dynamic engineering is the lack of public accessibility to dynamic data. An open-access protein dynamics database would accelerate the successful implementation of dynamic enzyme engineering and broaden its application to many systems. Just as structural data is organized in the Protein Data Bank¹¹¹ and the Electron Microscopy Data Bank¹¹², protein design and

engineering data is retrievable from ProtaBank¹¹³, kinetics data is accessible in BRENDA¹¹⁴ and functional data from multiplexed assays of variant effect is becoming available through MaveBD¹¹⁵. Similarly, we envisage the development of a curated repository for dynamic data. Links to those other databases would create a fully integrated environment for depositing and retrieving structural, functional and dynamic data on enzyme systems.

The computational community is discussing remote data viewing and sharing, and is beginning to build such data-sharing initiatives^{116,117}. This will not be trivial to implement, in part because dynamic data is resource intensive but mostly because dynamic knowledge is defined using a variety of metrics ranging as widely as atomic relaxation metrics in NMR (such as R_1 , R_2 and $^{15}\text{N}\{^1\text{H}\}$ NOE to infer order parameters) to the crystallographic B-factor, RMSF derived from MD simulations and FRET frequencies. Including data acquired by varied experimental methods will be challenging, yet it is in the interest of the enzyme engineering community to harmonize and organize dynamic information into a standardized and retrievable format. As is the case for all publicly available, easy-to-use curated databases, funding to support continued development, data curation and database maintenance must be durably secured.

In conclusion, the feasibility and benefits of dynamic engineering have been demonstrated and the field of enzyme engineering is now clearly poised to systematically incorporate rational dynamic engineering in its pipeline toward high-performance engineered enzymes²². We envisage that the broad implementation of dynamic engineering of enzyme function will necessarily require a massive acceleration of data acquisition to relate enzyme function and dynamics. Then, just as ML has revolutionized the field of protein structure prediction with AlphaFold2¹¹⁸ and has been democratized by ColabFold¹¹⁹, the integration of dynamics in these artificial intelligence workflows will offer unprecedented opportunities to predict and integrate atomic-scale motions in enzyme engineering and design.

A1.6 Funding

This work was supported by operating grants RGPIN-2018- 04686 (to J.N.P.) and RGPIN-2022-04368 (to N.D.) from the Natural Science and Engineering Research Council of Canada (NSERC), and the Canada Research Chairs Program (to J.N.P.). C.L.S.D. was supported by a PhD graduate scholarship from Natural Science and Engineering Research Council of Canada (NSERC) and by Université de Montréal. N.D. holds a Research Scholar Senior Career Award (281993) from the Fonds de Recherche Québec—Santé (FRQS).

PEDS board member: Prof. Christopher Snow.

A1.7 Acknowledgments

The authors dedicate this review to the memory of Danny Tawfik, explorer of protein evolution, deep thinker and visionary.

A1.8 References

- (1) Xu, Z.; Cen, Y.-K.; Zou, S.-P.; Xue, Y.-P.; Zheng, Y.-G. Recent Advances in the Improvement of Enzyme Thermostability by Structure Modification. *Crit. Rev. Biotechnol.* **2020**, *40* (1), 83–98. <https://doi.org/10.1080/07388551.2019.1682963>.
- (2) Renata, H.; Wang, Z. J.; Arnold, F. H. Expanding the Enzyme Universe: Accessing Non-Natural Reactions by Mechanism-Guided Directed Evolution. *Angew. Chem. Int. Ed.* **2015**, *54* (11), 3351–3367. <https://doi.org/10.1002/anie.201409470>.
- (3) Coelho, P. S.; Brustad, E. M.; Kannan, A.; Arnold, F. H. Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes. *Science* **2013**, *339* (6117), 307–310. <https://doi.org/10.1126/science.1231434>.
- (4) McIntosh, J. A.; Farwell, C. C.; Arnold, F. H. Expanding P450 Catalytic Reaction Space through Evolution and Engineering. *Curr. Opin. Chem. Biol.* **2014**, *19*, 126–134. <https://doi.org/10.1016/j.cbpa.2014.02.001>.
- (5) Devine, P. N.; Howard, R. M.; Kumar, R.; Thompson, M. P.; Truppo, M. D.; Turner, N. J. Extending the Application of Biocatalysis to Meet the Challenges of Drug Development. *Nat. Rev. Chem.* **2018**, *2* (12), 409–421. <https://doi.org/10.1038/s41570-018-0055-1>.
- (6) Wu, S.; Snajdrova, R.; Moore, J. C.; Baldenius, K.; Bornscheuer, U. T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angew. Chem. Int. Ed.* **2021**, *60* (1), 88–119. <https://doi.org/10.1002/anie.202006648>.
- (7) Huffman, M. A.; Fryszkowska, A.; Alvizo, O.; Borra-Garske, M.; Campos, K. R.; Canada, K. A.; Devine, P. N.; Duan, D.; Forstater, J. H.; Grosser, S. T.; Halsey, H. M.; Hughes, G. J.; Jo, J.; Joyce, L. A.; Kolev, J. N.; Liang, J.; Maloney, K. M.; Mann, B. F.; Marshall, N. M.; McLaughlin, M.; Moore, J. C.; Murphy, G. S.; Nawrat, C. C.; Nazor, J.; Novick, S.; Patel, N. R.; Rodriguez-Granillo, A.; Robaire, S. A.; Sherer, E. C.; Truppo, M. D.; Whittaker, A. M.; Verma, D.; Xiao, L.; Xu, Y.; Yang, H. Design of an in Vitro Biocatalytic Cascade for the Manufacture of Islatravir. *Science* **2019**, *366* (6470), 1255–1259. <https://doi.org/10.1126/science.aay8484>.
- (8) Boehr, D. D.; D’Amico, R. N.; O’Rourke, K. F. Engineered Control of Enzyme Structural Dynamics and Function. *Protein Sci.* **2018**, *27* (4), 825–838. <https://doi.org/10.1002/pro.3379>.
- (9) Kovermann, M.; Rogne, P.; Wolf-Watz, M. Protein Dynamics and Function from Solution State NMR Spectroscopy. *Q. Rev. Biophys.* **2016**, *49*, e6. <https://doi.org/10.1017/S0033583516000019>.
- (10) Huang, P. S.; Boyken, S. E.; Baker, D. The Coming of Age of de Novo Protein Design. *Nature* **2016**, *537* (7620), 320–327. <https://doi.org/10.1038/nature19946>.
- (11) Woolfson, D. N. A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. *J. Mol. Biol.* **2021**, *433* (20), 167160. <https://doi.org/10.1016/j.jmb.2021.167160>.
- (12) Alejaldre, L.; Pelletier, J. N.; Quaglia, D. Methods for Enzyme Library Creation: Which One Will You Choose?: A Guide for Novices and Experts to Introduce Genetic Diversity. *BioEssays* **2021**, *43* (8), 2100052. <https://doi.org/10.1002/bies.202100052>.

- (13) Petrović, D.; Risso, V. A.; Kamerlin, S. C. L.; Sanchez-Ruiz, J. M. Conformational Dynamics and Enzyme Evolution. *J. R. Soc. Interface* **2018**, *15* (144), 20180330. <https://doi.org/10.1098/rsif.2018.0330>.
- (14) Wolf-Watz, M.; Thai, V.; Henzler-Wildman, K.; Hadjipavlou, G.; Eisenmesser, E. Z.; Kern, D. Linkage between Dynamics and Catalysis in a Thermophilic-Mesophilic Enzyme Pair. *Nat. Struct. Mol. Biol.* **2004**, *11* (10), 945–949. <https://doi.org/10.1038/nsmb821>.
- (15) Watt, E. D.; Shimada, H.; Kovrigin, E. L.; Loria, J. P. The Mechanism of Rate-Limiting Motions in Enzyme Function. *Proc. Natl. Acad. Sci.* **2007**, *104* (29), 11981–11986. <https://doi.org/10.1073/pnas.0702551104>.
- (16) Gora, A.; Brezovsky, J.; Damborsky, J. Gates of Enzymes. *Chem. Rev.* **2013**, *113* (8), 5871–5923. <https://doi.org/10.1021/cr300384w>.
- (17) Gardner, J. M.; Biler, M.; Risso, V. A.; Sanchez-Ruiz, J. M.; Kamerlin, S. C. L. Manipulating Conformational Dynamics to Repurpose Ancient Proteins for Modern Catalytic Functions. *ACS Catal* **2020**, *10* (9), 4863–4870. <https://doi.org/10.1021/acscatal.0c00722>.
- (18) Berg, J. M.; Tymoczko, J. L.; Stryer, L.; Clarke, N. D. *Biochemistry, 5th Edition, Chapter 8 Enzymes: Basic Concepts and Kinetics*, 5. ed., [Nachdr.], international ed.; Freeman: New York, 2003.
- (19) Smejkal, G. B.; Kakumanu, S. Enzymes and Their Turnover Numbers. *Expert Rev. Proteomics* **2019**, *16* (7), 543–544. <https://doi.org/10.1080/14789450.2019.1630275>.
- (20) Clausen, R.; Shehu, A. A Data-Driven Evolutionary Algorithm for Mapping Multibasin Protein Energy Landscapes. *J. Comput. Biol.* **2015**, *22* (9), 844–860. <https://doi.org/10.1089/cmb.2015.0107>.
- (21) Maria-Solano, M. A.; Serrano-Hervás, E.; Romero-Rivera, A.; Iglesias-Fernández, J.; Osuna, S. Role of Conformational Dynamics in the Evolution of Novel Enzyme Function. *Chem. Commun.* **2018**, *54* (50), 6622–6634. <https://doi.org/10.1039/C8CC02426J>.
- (22) Damry, A. M.; Jackson, C. J. The Evolution and Engineering of Enzyme Activity through Tuning Conformational Landscapes. *Protein Eng. Des. Sel.* **2021**, *34*, gzab009. <https://doi.org/10.1093/protein/gzab009>.
- (23) Pabis, A.; Risso, V. A.; Sanchez-Ruiz, J. M.; Kamerlin, S. C. Cooperativity and Flexibility in Enzyme Evolution. *Curr Opin Struct Biol* **2018**, *48*, 83–92. <https://doi.org/10.1016/j.sbi.2017.10.020>.
- (24) Kaczmarek, J. A.; Mahawaththa, M. C.; Feintuch, A.; Clifton, B. E.; Adams, L. A.; Goldfarb, D.; Otting, G.; Jackson, C. J. Altered Conformational Sampling along an Evolutionary Trajectory Changes the Catalytic Activity of an Enzyme. *Nat. Commun.* **2020**, *11* (1), 5945. <https://doi.org/10.1038/s41467-020-19695-9>.
- (25) Romero-Rivera, A.; Garcia-Borràs, M.; Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **2017**, *7* (12), 8524–8532. <https://doi.org/10.1021/acscatal.7b02954>.

- (26) Narayanan, C.; Bernard, D. N.; Bafna, K.; Gagné, D.; Chennubhotla, C. S.; Doucet, N.; Agarwal, P. K. Conservation of Dynamics Associated with Biological Function in an Enzyme Superfamily. *Structure* **2018**, *26* (3), 426–436.e3. <https://doi.org/10.1016/j.str.2018.01.015>.
- (27) Bhabha, G.; Ekiert, D. C.; Jennewein, M.; Zmasek, C. M.; Tuttle, L. M.; Kroon, G.; Dyson, H. J.; Godzik, A.; Wilson, I. A.; Wright, P. E. Divergent Evolution of Protein Conformational Dynamics in Dihydrofolate Reductase. *Nat. Struct. Mol. Biol.* **2013**, *20* (11), 1243–1249. <https://doi.org/10.1038/nsmb.2676>.
- (28) Nevin Gerek, Z.; Kumar, S.; Banu Ozkan, S. Structural Dynamics Flexibility Informs Function and Evolution at a Proteome Scale. *Evol. Appl.* **2013**, *6* (3), 423–433. <https://doi.org/10.1111/eva.12052>.
- (29) Klinman, J. P.; Kohen, A. Evolutionary Aspects of Enzyme Dynamics. *J. Biol. Chem.* **2014**, *289* (44), 30205–30212. <https://doi.org/10.1074/jbc.R114.565515>.
- (30) Saavedra, H. G.; Wrabl, J. O.; Anderson, J. A.; Li, J.; Hilser, V. J. Dynamic Allostery Can Drive Cold Adaptation in Enzymes. *Nature* **2018**, *558* (7709), 324–328. <https://doi.org/10.1038/s41586-018-0183-2>.
- (31) Campbell, E.; Kaltenbach, M.; Correy, G. J.; Carr, P. D.; Porebski, B. T.; Livingstone, E. K.; Afriat-Jurnou, L.; Buckle, A. M.; Weik, M.; Hollfelder, F.; Tokuriki, N.; Jackson, C. J. The Role of Protein Dynamics in the Evolution of New Enzyme Function. *Nat. Chem. Biol.* **2016**, *12* (11), 944–950. <https://doi.org/10.1038/nchembio.2175>.
- (32) Campbell, E. C.; Correy, G. J.; Mabbitt, P. D.; Buckle, A. M.; Tokuriki, N.; Jackson, C. J. Laboratory Evolution of Protein Conformational Dynamics. *Curr. Opin. Struct. Biol.* **2018**, *50*, 49–57. <https://doi.org/10.1016/j.sbi.2017.09.005>.
- (33) Bunzel, H. A.; Anderson, J. L. R.; Hilvert, D.; Arcus, V. L.; van der Kamp, M. W.; Mulholland, A. J. Evolution of Dynamical Networks Enhances Catalysis in a Designer Enzyme. *Nat. Chem.* **2021**, *13* (10), 1017–1022. <https://doi.org/10.1038/s41557-021-00763-6>.
- (34) Gobeil, S. M. C.; Ebert, M. C. C. J. C.; Park, J.; Gagné, D.; Doucet, N.; Berghuis, A. M.; Pleiss, J.; Pelletier, J. N. The Structural Dynamics of Engineered β -Lactamases Vary Broadly on Three Timescales yet Sustain Native Function. *Sci. Rep.* **2019**, *9* (1), 6656. <https://doi.org/10.1038/s41598-019-42866-8>.
- (35) Francis, K.; Stojković, V.; Kohen, A. Preservation of Protein Dynamics in Dihydrofolate Reductase Evolution. *J. Biol. Chem.* **2013**, *288* (50), 35961–35968. <https://doi.org/10.1074/jbc.M113.507632>.
- (36) Liu, C. T.; Hanoian, P.; French, J. B.; Pringle, T. H.; Hammes-Schiffer, S.; Benkovic, S. J. Functional Significance of Evolving Protein Sequence in Dihydrofolate Reductase from Bacteria to Humans. *Proc. Natl. Acad. Sci.* **2013**, *110* (25), 10159–10164. <https://doi.org/10.1073/pnas.1307130110>.
- (37) Singh, A.; Fenwick, R. B.; Dyson, H. J.; Wright, P. E. Role of Active Site Loop Dynamics in Mediating Ligand Release from *E. Coli* Dihydrofolate Reductase. *Biochemistry* **2021**, *60* (35), 2663–2671. <https://doi.org/10.1021/acs.biochem.1c00461>.
- (38) van den Bedem, H.; Fraser, J. S. Integrative, Dynamic Structural Biology at Atomic Resolution—It’s about Time. *Nat. Methods* **2015**, *12* (4), 307–318. <https://doi.org/10.1038/nmeth.3324>.

- (39) Agarwal, P. K.; Bernard, D. N.; Bafna, K.; Doucet, N. Enzyme Dynamics: Looking Beyond a Single Structure. *ChemCatChem* **2020**, *12* (19), 4704–4720. <https://doi.org/10.1002/cctc.202000665>.
- (40) Srivastava, A.; Nagai, T.; Srivastava, A.; Miyashita, O.; Tama, F. Role of Computational Methods in Going beyond X-Ray Crystallography to Explore Protein Structure and Dynamics. *Int. J. Mol. Sci.* **2018**, *19* (11), 3401. <https://doi.org/10.3390/ijms19113401>.
- (41) Broom, A.; Rakotoharisoa, R. V.; Thompson, M. C.; Zarifi, N.; Nguyen, E.; Mukhametzhanov, N.; Liu, L.; Fraser, J. S.; Chica, R. A. Ensemble-Based Enzyme Design Can Recapitulate the Effects of Laboratory Directed Evolution in Silico. *Nat. Commun.* **2020**, *11* (1), 4808. <https://doi.org/10.1038/s41467-020-18619-x>.
- (42) Fischer, M. Macromolecular Room Temperature Crystallography. *Q. Rev. Biophys.* **2021**, *54*, e1. <https://doi.org/10.1017/S0033583520000128>.
- (43) Martin-Garcia, J. M. Protein Dynamics and Time Resolved Protein Crystallography at Synchrotron Radiation Sources: Past, Present and Future. *Crystals* **2021**, *11* (5), 521. <https://doi.org/10.3390/cryst11050521>.
- (44) Knox, R.; Lento, C.; Wilson, D. J. Mapping Conformational Dynamics to Individual Steps in the TEM-1 β -Lactamase Catalytic Mechanism. *J. Mol. Biol.* **2018**, *430* (18), 3311–3322. <https://doi.org/10.1016/j.jmb.2018.06.045>.
- (45) Mazal, H.; Haran, G. Single-Molecule FRET Methods to Study the Dynamics of Proteins at Work. *Curr. Opin. Biomed. Eng.* **2019**, *12*, 8–17. <https://doi.org/10.1016/j.cobme.2019.08.007>.
- (46) Narang, D.; Lento, C.; J. Wilson, D. HDX-MS: An Analytical Tool to Capture Protein Motion in Action. *Biomedicines* **2020**, *8* (7), 224. <https://doi.org/10.3390/biomedicines8070224>.
- (47) Matsumoto, S.; Ishida, S.; Araki, M.; Kato, T.; Terayama, K.; Okuno, Y. Extraction of Protein Dynamics Information from Cryo-EM Maps Using Deep Learning. *Nat. Mach. Intell.* **2021**, *3* (2), 153–160. <https://doi.org/10.1038/s42256-020-00290-y>.
- (48) Lento, C.; Wilson, D. J. Subsecond Time-Resolved Mass Spectrometry in Dynamic Structural Biology. *Chem. Rev.* **2022**, *122* (8), 7624–7646. <https://doi.org/10.1021/acs.chemrev.1c00222>.
- (49) Tsai, M.-D.; Wu, W.-J.; Ho, M.-C. Enzymology and Dynamics by Cryogenic Electron Microscopy. *Annu. Rev. Biophys.* **2022**, *51* (1), 19–38. <https://doi.org/10.1146/annurev-biophys-100121-075228>.
- (50) Konovalov, K. A.; Unarta, I. C.; Cao, S.; Goonetilleke, E. C.; Huang, X. Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning. *JACS Au* **2021**, *1* (9), 1330–1341. <https://doi.org/10.1021/jacsau.1c00254>.
- (51) Atilgan, A. R.; Atilgan, C. Computational Strategies for Protein Conformational Ensemble Detection. *Curr. Opin. Struct. Biol.* **2022**, *72*, 79–87. <https://doi.org/10.1016/j.sbi.2021.08.007>.
- (52) Hospital, A.; Andrio, P.; Fenollosa, C.; Cicin-Sain, D.; Orozco, M.; Gelpí, J. L. MDWeb and MDMoby: An Integrated Web-Based Platform for Molecular Dynamics Simulations. *Bioinformatics* **2012**, *28* (9), 1278–1279. <https://doi.org/10.1093/bioinformatics/bts139>.

- (53) Madadkar-Sobhani, A.; Guallar, V. PELE Web Server: Atomistic Study of Biomolecular Systems at Your Fingertips. *Nucleic Acids Res* **2013**, *41* (Web Server issue), W322-8. <https://doi.org/10.1093/nar/gkt454>.
- (54) Marchetto, A.; Si Chaib, Z.; Rossi, C. A.; Ribeiro, R.; Pantano, S.; Rossetti, G.; Giorgetti, A. CGMD Platform: Integrated Web Servers for the Preparation, Running, and Analysis of Coarse-Grained Molecular Dynamics Simulations. *Molecules* **2020**, *25* (24), 5934. <https://doi.org/10.3390/molecules25245934>.
- (55) Gorman, S. D.; D'Amico, R. N.; Winston, D. S.; Boehr, D. D. Engineering Allostery into Proteins. In *Protein Allostery in Drug Discovery*; Zhang, J., Nussinov, R., Eds.; Advances in Experimental Medicine and Biology; Springer Singapore: Singapore, 2019; Vol. 1163, pp 359–384. https://doi.org/10.1007/978-981-13-8719-7_15.
- (56) Kokkonen, P.; Bednar, D.; Pinto, G.; Prokop, Z.; Damborsky, J. Engineering Enzyme Access Tunnels. *Biotechnol. Adv.* **2019**, *37* (6), 107386. <https://doi.org/10.1016/j.biotechadv.2019.04.008>.
- (57) Liu, Y.; Xu, G.; Zhou, J.; Ni, J.; Zhang, L.; Hou, X.; Yin, D.; Rao, Y.; Zhao, Y.-L.; Ni, Y. Structure-Guided Engineering of d-Carbamoylase Reveals a Key Loop at Substrate Entrance Tunnel. *ACS Catal.* **2020**, *10* (21), 12393–12402. <https://doi.org/10.1021/acscatal.0c02942>.
- (58) Planas-Iglesias, J.; Opaleny, F.; Ulbrich, P.; Stourac, J.; Sanusi, Z.; Pinto, G. P.; Schenkmyerova, A.; Byska, J.; Damborsky, J.; Kozlikova, B.; Bednar, D. LoopGrafter: A Web Tool for Transplanting Dynamical Loops for Protein Engineering. *Nucleic Acids Res.* **2022**, gkac249. <https://doi.org/10.1093/nar/gkac249>.
- (59) Chen, X.; Schwartz, S. D. Directed Evolution as a Probe of Rate Promoting Vibrations Introduced via Mutational Change. *Biochemistry* **2018**, *57* (23), 3289–3298. <https://doi.org/10.1021/acs.biochem.8b00185>.
- (60) Khersonsky, O.; Kiss, G.; Röthlisberger, D.; Dym, O.; Albeck, S.; Houk, K. N.; Baker, D.; Tawfik, D. S. Bridging the Gaps in Design Methodologies by Evolutionary Optimization of the Stability and Proficiency of Designed Kemp Eliminase KE59. *Proc. Natl. Acad. Sci.* **2012**, *109* (26), 10358–10363. <https://doi.org/10.1073/pnas.1121063109>.
- (61) Pudney, C. R.; Guerriero, A.; Baxter, N. J.; Johannissen, L. O.; Waltho, J. P.; Hay, S.; Scrutton, N. S. Fast Protein Motions Are Coupled to Enzyme H-Transfer Reactions. *J. Am. Chem. Soc.* **2013**, *135* (7), 2512–2517. <https://doi.org/10.1021/ja311277k>.
- (62) Chen, X.; Schwartz, S. D. Multiple Reaction Pathways in the Morphinone Reductase-Catalyzed Hydride Transfer Reaction. *ACS Omega* **2020**, *5* (36), 23468–23480. <https://doi.org/10.1021/acsomega.0c03472>.
- (63) Lehmann, M.; Pasamontes, L.; Lassen, S. F.; Wyss, M. The Consensus Concept for Thermostability Engineering of Proteins. *Biochim. Biophys. Acta BBA - Protein Struct. Mol. Enzymol.* **2000**, *1543* (2), 408–415. [https://doi.org/10.1016/S0167-4838\(00\)00238-7](https://doi.org/10.1016/S0167-4838(00)00238-7).
- (64) Bai, X.; Li, D.; Ma, F.; Deng, X.; Luo, M.; Feng, Y.; Yang, G. Improved Thermostability of Creatinase from *Alcaligenes Faecalis* through Non-Biased Phylogenetic Consensus-Guided

- Mutagenesis. *Microb. Cell Factories* **2020**, *19* (1), 194. <https://doi.org/10.1186/s12934-020-01451-9>.
- (65) Goldenzweig, A.; Goldsmith, M.; Hill, S. E.; Gertman, O.; Laurino, P.; Ashani, Y.; Dym, O.; Unger, T.; Albeck, S.; Prilusky, J.; Lieberman, R. L.; Aharoni, A.; Silman, I.; Sussman, J. L.; Tawfik, D. S.; Fleishman, S. J. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **2016**, *63* (2), 337–346. <https://doi.org/10.1016/j.molcel.2016.06.012>.
- (66) Hettiaratchi, M. H.; O’Meara, M. J.; O’Meara, T. R.; Pickering, A. J.; Letko-Khait, N.; Shoichet, M. S. Reengineering Biocatalysts: Computational Redesign of Chondroitinase ABC Improves Efficacy and Stability. *Sci. Adv.* **2020**, *6* (34), eabc6378. <https://doi.org/10.1126/sciadv.abc6378>.
- (67) Doble, M. V.; Obrecht, L.; Joosten, H.-J.; Lee, M.; Rozeboom, H. J.; Branigan, E.; Naismith, James. H.; Janssen, D. B.; Jarvis, A. G.; Kamer, P. C. J. Engineering Thermostability in Artificial Metalloenzymes to Increase Catalytic Activity. *ACS Catal.* **2021**, *11* (6), 3620–3627. <https://doi.org/10.1021/acscatal.0c05413>.
- (68) Markova, K.; Kunka, A.; Chmelova, K.; Havlasek, M.; Babkova, P.; Marques, S. M.; Vasina, M.; Planas-Iglesias, J.; Chaloupkova, R.; Bednar, D.; Prokop, Z.; Damborsky, J.; Marek, M. Computational Enzyme Stabilization Can Affect Folding Energy Landscapes and Lead to Catalytically Enhanced Domain-Swapped Dimers. *ACS Catal.* **2021**, *11* (21), 12864–12885. <https://doi.org/10.1021/acscatal.1c03343>.
- (69) Stourac, J.; Dubrava, J.; Musil, M.; Horackova, J.; Damborsky, J.; Mazurenko, S.; Bednar, D. FireProtDB: Database of Manually Curated Protein Stability Data. *Nucleic Acids Res.* **2021**, *49* (D1), D319–D324. <https://doi.org/10.1093/nar/gkaa981>.
- (70) Gonzalez, N. A.; Li, B. A.; McCully, M. E. The Stability and Dynamics of Computationally Designed Proteins. *Protein Eng. Des. Sel.* **2022**, *35*, gzac001. <https://doi.org/10.1093/protein/gzac001>.
- (71) Parra-Cruz, R.; Jäger, C. M.; Lau, P. L.; Gomes, R. L.; Pordea, A. Rational Design of Thermostable Carbonic Anhydrase Mutants Using Molecular Dynamics Simulations. *J. Phys. Chem. B* **2018**, *122* (36), 8526–8536. <https://doi.org/10.1021/acs.jpcc.8b05926>.
- (72) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. *Nucleic Acids Res.* **2005**, *33* (Web Server), W382–W388. <https://doi.org/10.1093/nar/gki387>.
- (73) Broom, A.; Trainor, K.; Jacobi, Z.; Meiering, E. M. Computational Modeling of Protein Stability: Quantitative Analysis Reveals Solutions to Pervasive Problems. *Structure* **2020**, *28* (6), 717–726.e3. <https://doi.org/10.1016/j.str.2020.04.003>.
- (74) Butterwick, J. A.; Patrick Loria, J.; Astrof, N. S.; Kroenke, C. D.; Cole, R.; Rance, M.; Palmer, A. G. Multiple Time Scale Backbone Dynamics of Homologous Thermophilic and Mesophilic Ribonuclease HI Enzymes. *J. Mol. Biol.* **2004**, *339* (4), 855–871. <https://doi.org/10.1016/j.jmb.2004.03.055>.

- (75) Karshikoff, A.; Nilsson, L.; Ladenstein, R. Rigidity versus Flexibility: The Dilemma of Understanding Protein Thermal Stability. *FEBS J.* **2015**, *282* (20), 3899–3917. <https://doi.org/10.1111/febs.13343>.
- (76) Barik, S. Evolution of Protein Structure and Stability in Global Warming. *Int. J. Mol. Sci.* **2020**, *21* (24), 9662. <https://doi.org/10.3390/ijms21249662>.
- (77) Ebert, M. C. C. J. C.; Guzman Espinola, J.; Lamoureux, G.; Pelletier, J. N. Substrate-Specific Screening for Mutational Hotspots Using Biased Molecular Dynamics Simulations. *ACS Catal.* **2017**, *7* (10), 6786–6797. <https://doi.org/10.1021/acscatal.7b02634>.
- (78) Rouhani, M.; Khodabakhsh, F.; Norouzian, D.; Cohan, R. A.; Valizadeh, V. Molecular Dynamics Simulation for Rational Protein Engineering: Present and Future Prospectus. *J. Mol. Graph. Model.* **2018**, *84*, 43–53. <https://doi.org/10.1016/j.jmgm.2018.06.009>.
- (79) Schenk Mayerova, A.; Pinto, G. P.; Toul, M.; Marek, M.; Hernychova, L.; Planas-Iglesias, J.; Daniel Liskova, V.; Pluskal, D.; Vasina, M.; Emond, S.; Dörr, M.; Chaloupkova, R.; Bednar, D.; Prokop, Z.; Hollfelder, F.; Bornscheuer, U. T.; Damborsky, J. Engineering the Protein Dynamics of an Ancestral Luciferase. *Nat. Commun.* **2021**, *12* (1), 3616. <https://doi.org/10.1038/s41467-021-23450-z>.
- (80) Bata, Z.; Molnár, Z.; Madaras, E.; Molnár, B.; Santa-Bell, E.; Varga, A.; Leveles, I.; Qian, R.; Hammerschmidt, F.; Paizs, C.; Vértessy, B. G.; Poppe, L. Substrate Tunnel Engineering Aided by X-Ray Crystallography and Functional Dynamics Swaps the Function of MIO-Enzymes. *ACS Catal.* **2021**, *11* (8), 4538–4549. <https://doi.org/10.1021/acscatal.1c00266>.
- (81) Spence, M. A.; Kaczmarek, J. A.; Saunders, J. W.; Jackson, C. J. Ancestral Sequence Reconstruction for Protein Engineers. *Curr. Opin. Struct. Biol.* **2021**, *69*, 131–141. <https://doi.org/10.1016/j.sbi.2021.04.001>.
- (82) Busch, F.; Rajendran, C.; Heyn, K.; Schlee, S.; Merkl, R.; Sterner, R. Ancestral Tryptophan Synthase Reveals Functional Sophistication of Primordial Enzyme Complexes. *Cell Chem. Biol.* **2016**, *23* (6), 709–715. <https://doi.org/10.1016/j.chembiol.2016.05.009>.
- (83) Maria-Solano, M. A.; Kinader, T.; Iglesias-Fernández, J.; Sterner, R.; Osuna, S. *In Silico* Identification and Experimental Validation of Distal Activity-Enhancing Mutations in Tryptophan Synthase. *ACS Catal.* **2021**, *11* (21), 13733–13743. <https://doi.org/10.1021/acscatal.1c03950>.
- (84) Zoi, I.; Suarez, J.; Antoniou, D.; Cameron, S. A.; Schramm, V. L.; Schwartz, S. D. Modulating Enzyme Catalysis through Mutations Designed to Alter Rapid Protein Dynamics. *J. Am. Chem. Soc.* **2016**, *138* (10), 3403–3409. <https://doi.org/10.1021/jacs.5b12551>.
- (85) Contreras-Llano, L. E.; Tan, C. High-Throughput Screening of Biomolecules Using Cell-Free Gene Expression Systems. *Synth. Biol.* **2018**, *3* (1), ysy012. <https://doi.org/10.1093/synbio/ysy012>.
- (86) Griffiths, A. D. Directed Evolution of an Extremely Fast Phosphotriesterase by *in Vitro* Compartmentalization. *EMBO J.* **2003**, *22* (1), 24–35. <https://doi.org/10.1093/emboj/cdg014>.
- (87) Bouzetos, E.; Ganar, K. A.; Mastrobattista, E.; Deshpande, S.; van der Oost, J. (R)Evolution-on-a-Chip. *Trends Biotechnol.* **2022**, *40* (1), 60–76. <https://doi.org/10.1016/j.tibtech.2021.04.009>.

- (88) Diefenbach, X. W.; Farasat, I.; Guetschow, E. D.; Welch, C. J.; Kennedy, R. T.; Sun, S.; Moore, J. C. Enabling Biocatalysis by High-Throughput Protein Engineering Using Droplet Microfluidics Coupled to Mass Spectrometry. *ACS Omega* **2018**, *3* (2), 1498–1508. <https://doi.org/10.1021/acsomega.7b01973>.
- (89) Tan, Y.; Zhang, Y.; Han, Y.; Liu, H.; Chen, H.; Ma, F.; Withers, S. G.; Feng, Y.; Yang, G. Directed Evolution of an A1,3-Fucosyltransferase Using a Single-Cell Ultrahigh-Throughput Screening Method. *Sci. Adv.* **2019**, *5* (10), eaaw8451. <https://doi.org/10.1126/sciadv.aaw8451>.
- (90) Yeom, S.-J.; Kim, M.; Kwon, K. K.; Fu, Y.; Rha, E.; Park, S.-H.; Lee, H.; Kim, H.; Lee, D.-H.; Kim, D.-M.; Lee, S.-G. A Synthetic Microbial Biosensor for High-Throughput Screening of Lactam Biocatalysts. *Nat. Commun.* **2018**, *9* (1), 5053. <https://doi.org/10.1038/s41467-018-07488-0>.
- (91) Armetta, J.; Berthome, R.; Cros, A.; Pophillat, C.; Colombo, B. M.; Pandi, A.; Grigoras, I. Biosensor-Based Enzyme Engineering Approach Applied to Psicose Biosynthesis. *Synth. Biol.* **2019**, *4* (1), ysz028. <https://doi.org/10.1093/synbio/ysz028>.
- (92) Kourist, R.; Brundiek, H.; Bornscheuer, U. T. Protein Engineering and Discovery of Lipases. *Eur. J. Lipid Sci. Technol.* **2010**, *112* (1), 64–74. <https://doi.org/10.1002/ejlt.200900143>.
- (93) Datta, S.; Rajnish, K. N.; Samuel, M. S.; Pugazlendhi, A.; Selvarajan, E. Metagenomic Applications in Microbial Diversity, Bioremediation, Pollution Monitoring, Enzyme and Drug Discovery. A Review. *Environ. Chem. Lett.* **2020**, *18* (4), 1229–1241. <https://doi.org/10.1007/s10311-020-01010-z>.
- (94) Robinson, S. L.; Piel, J.; Sunagawa, S. A Roadmap for Metagenomic Enzyme Discovery. *Nat. Prod. Rep.* **2021**, *38* (11), 1994–2023. <https://doi.org/10.1039/D1NP00006C>.
- (95) Luk, L. Y. P.; Javier Ruiz-Pernía, J.; Dawson, W. M.; Roca, M.; Loveridge, E. J.; Glowacki, D. R.; Harvey, J. N.; Mulholland, A. J.; Tuñón, I.; Moliner, V.; Allemann, R. K. Unraveling the Role of Protein Dynamics in Dihydrofolate Reductase Catalysis. *Proc. Natl. Acad. Sci.* **2013**, *110* (41), 16344–16349. <https://doi.org/10.1073/pnas.1312437110>.
- (96) Li, D.; Liu, M. S.; Ji, B. Mapping the Dynamics Landscape of Conformational Transitions in Enzyme: The Adenylate Kinase Case. *Biophys. J.* **2015**, *109* (3), 647–660. <https://doi.org/10.1016/j.bpj.2015.06.059>.
- (97) Alderson, T. R.; Kay, L. E. NMR Spectroscopy Captures the Essential Role of Dynamics in Regulating Biomolecular Function. *Cell* **2021**, *184* (3), 577–595. <https://doi.org/10.1016/j.cell.2020.12.034>.
- (98) Surpeta, B.; Sequeiros-Borja, C.; Brezovsky, J. Dynamics, a Powerful Component of Current and Future in Silico Approaches for Protein Design and Engineering. *Int. J. Mol. Sci.* **2020**, *21* (8), 2713. <https://doi.org/10.3390/ijms21082713>.
- (99) Marques, S. M.; Planas-Iglesias, J.; Damborsky, J. Web-Based Tools for Computational Enzyme Design. *Curr. Opin. Struct. Biol.* **2021**, *69*, 19–34. <https://doi.org/10.1016/j.sbi.2021.01.010>.

- (100) Sequeiros-Borja, C. E.; Surpeta, B.; Brezovsky, J. Recent Advances in User-Friendly Computational Tools to Engineer Protein Function. *Brief. Bioinform.* **2021**, *22* (3), bbaa150. <https://doi.org/10.1093/bib/bbaa150>.
- (101) David, C. C.; Jacobs, D. J. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. In *Protein Dynamics*; Livesay, D. R., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2014; Vol. 1084, pp 193–226. https://doi.org/10.1007/978-1-62703-658-0_11.
- (102) Siedhoff, N. E.; Schwaneberg, U.; Davari, M. D. Machine Learning-Assisted Enzyme Engineering. In *Methods in Enzymology*; Elsevier, 2020; Vol. 643, pp 281–315. <https://doi.org/10.1016/bs.mie.2020.05.005>.
- (103) Ferguson, A. L.; Ranganathan, R. 100th Anniversary of Macromolecular Science Viewpoint: Data-Driven Protein Design. *ACS Macro Lett.* **2021**, *10* (3), 327–340. <https://doi.org/10.1021/acsmacrolett.0c00885>.
- (104) Rudden, L. S. P.; Hijazi, M.; Barth, P. Deep Learning Approaches for Conformational Flexibility and Switching Properties in Protein Design. *Front. Mol. Biosci.* **2022**, *9*, 928534. <https://doi.org/10.3389/fmolb.2022.928534>.
- (105) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020**, *10* (2), 1210–1223. <https://doi.org/10.1021/acscatal.9b04321>.
- (106) Barros, E. P.; Schiffer, J. M.; Vorobieva, A.; Dou, J.; Baker, D.; Amaro, R. E. Improving the Efficiency of Ligand-Binding Protein Design with Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2019**, *15* (10), 5703–5715. <https://doi.org/10.1021/acs.jctc.9b00483>.
- (107) Lovelock, S. L.; Crawshaw, R.; Basler, S.; Levy, C.; Baker, D.; Hilvert, D.; Green, A. P. The Road to Fully Programmable Protein Catalysis. *Nature* **2022**, *606* (7912), 49–58. <https://doi.org/10.1038/s41586-022-04456-z>.
- (108) Otten, R.; Pádua, R. A. P.; Bunzel, H. A.; Nguyen, V.; Pitsawong, W.; Patterson, M.; Sui, S.; Perry, S. L.; Cohen, A. E.; Hilvert, D.; Kern, D. How Directed Evolution Reshapes the Energy Landscape in an Enzyme to Boost Catalysis. *Science* **2020**, *370* (6523), 1442–1446. <https://doi.org/10.1126/science.abd3623>.
- (109) Khersonsky, O.; Röthlisberger, D.; Wollacott, A. M.; Murphy, P.; Dym, O.; Albeck, S.; Kiss, G.; Houk, K. N.; Baker, D.; Tawfik, D. S. Optimization of the In-Silico-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution. *J. Mol. Biol.* **2011**, *407* (3), 391–412. <https://doi.org/10.1016/j.jmb.2011.01.041>.
- (110) Vaissier Welborn, V.; Head-Gordon, T. Computational Design of Synthetic Enzymes. *Chem. Rev.* **2019**, *119* (11), 6613–6630. <https://doi.org/10.1021/acs.chemrev.8b00399>.
- (111) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. <https://doi.org/10.1093/nar/28.1.235>.
- (112) Lawson, C. L.; Patwardhan, A.; Baker, M. L.; Hryc, C.; Garcia, E. S.; Hudson, B. P.; Lagerstedt, I.; Ludtke, S. J.; Pintilie, G.; Sala, R.; Westbrook, J. D.; Berman, H. M.; Kleywegt, G. J.; Chiu, W.

- EMDataBank Unified Data Resource for 3DEM. *Nucleic Acids Res.* **2016**, *44* (D1), D396–D403. <https://doi.org/10.1093/nar/gkv1126>.
- (113) Wang, C. Y.; Chang, P. M.; Ary, M. L.; Allen, B. D.; Chica, R. A.; Mayo, S. L.; Olafson, B. D. ProtaBank: A Repository for Protein Design and Engineering Data: ProtaBank: A Protein Engineering Database. *Protein Sci.* **2018**, *27* (6), 1113–1124. <https://doi.org/10.1002/pro.3406>.
- (114) Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblitz, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D. BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates. *Nucleic Acids Res.* **2021**, *49* (D1), D498–D508. <https://doi.org/10.1093/nar/gkaa1025>.
- (115) Esposito, D.; Weile, J.; Shendure, J.; Starita, L. M.; Papenfuss, A. T.; Roth, F. P.; Fowler, D. M.; Rubin, A. F. MaveDB: An Open-Source Platform to Distribute and Interpret Data from Multiplexed Assays of Variant Effect. *Genome Biol.* **2019**, *20* (1), 223. <https://doi.org/10.1186/s13059-019-1845-6>.
- (116) Abraham, M.; Apostolov, R.; Barnoud, J.; Bauer, P.; Blau, C.; Bonvin, A. M. J. J.; Chavent, M.; Chodera, J.; Čondić-Jurkić, K.; Delemotte, L.; Grubmüller, H.; Howard, R. J.; Jordan, E. J.; Lindahl, E.; Ollila, O. H. S.; Selent, J.; Smith, D. G. A.; Stansfeld, P. J.; Tiemann, J. K. S.; Trellet, M.; Woods, C.; Zhmurov, A. Sharing Data from Molecular Simulations. *J. Chem. Inf. Model.* **2019**, *59* (10), 4093–4099. <https://doi.org/10.1021/acs.jcim.9b00665>.
- (117) Kampfrath, M.; Staritzbichler, R.; Hernández, G. P.; Rose, A. S.; Tiemann, J. K. S.; Scheuermann, G.; Wiegrefe, D.; Hildebrand, P. W. MDsrv: Visual Sharing and Analysis of Molecular Dynamics Simulations. *Nucleic Acids Res.* **2022**, gkac398. <https://doi.org/10.1093/nar/gkac398>.
- (118) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- (119) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat. Methods* **2022**, *19* (6), 679–682. <https://doi.org/10.1038/s41592-022-01488-1>.
- (120) Pham, N. T. H.; Létourneau, M.; Fortier, M.; Bégin, G.; Al-Abdul-Wahid, M. S.; Pucci, F.; Folch, B.; Rومان, M.; Chatenet, D.; St-Pierre, Y.; Lagüe, P.; Calmettes, C.; Doucet, N. Perturbing Dimer Interactions and Allosteric Communication Modulates the Immunosuppressive Activity of Human Galectin-7. *J. Biol. Chem.* **2021**, *297* (5), 101308. <https://doi.org/10.1016/j.jbc.2021.101308>.