Characterisation, Classification and Conformational Variability of Organic Enzyme Cofactors

Julia D. Fischer

European Bioinformatics Institute
Clare Hall College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

11 April 2011

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the word limit of 60,000 words.

Acknowledgements

I would like to thank all the members of the Thornton research group for their constant interest in my work, their continuous willingness to answer my academic questions, and for their company during my time at the EBI. This includes Saumya Kumar, Sergio Martinez Cuesta, Matthias Ziehm, Dr. Daniela Wieser, Dr. Xun Li, Dr. Irene Papatheodorou, Dr. Pedro Ballester, Dr. Abdullah Kahraman, Dr. Rafael Najmanovich, Dr. Tjaart de Beer, Dr. Syed Asad Rahman, Dr. Nicholas Furnham, Dr. Roman Laskowski and Dr. Gemma Holliday. Special thanks to Asad for allowing me to use early development versions of his SMSD software and for help and advice with the KEGG API installation, to Roman for knowing where to find all kinds of data, to Dani for help with R scripts, to Nick for letting me use his E.C. tree program, to Tjaart for python advice and especially to Gemma for her constant advice and feedback on my work in all aspects, in particular the chemistry side. Most importantly, I would like to thank Prof. Janet Thornton for giving me the chance to work on this project, for all the time she spent in meetings with me and reading my work, for sharing her seemingly limitless knowledge and enthusiasm about the fascinating world of enzymes, and for being such an experienced and motivational advisor. I would also like to thank my thesis advisory committee: Prof. Christoph Steinbeck, Prof. John Mitchell and Prof. Victor Lamzin for accompanying my scientific journey, for their continuous critical evaluation of my work, and for their scientific as well as organisational advice. Further, I would like to thank Helen Barker and Stacy Schab who were very helpful in managing my meetings with Janet and my travelling needs.

For proofreading, I would like to thank Dr. Steve Wilder, Dr. Gilleain Torrance, Sergio Martinez Cuesta, Dr. Nick Furnham, Dr. Roman Laskowski, Dr. Cajetan Neubauer and of course Janet and Gemma. Most of the proofreading was however done by Dr. Tjaart de Beer who deserves special thanks.

My time at the EBI would not have been the same without the company of my fellow predocs. I would like to thank all of them for the welcome distractions and - when necessary - their commiseration at the weekly predoc lunches and at the many social activities. I would like to especially thank Judith Zaugg and Michele Mattioni, with whom I shared the predoc experience from the very beginning.

In hindsight I could not imagine my life here without rowing. I would like to thank Dr. Adrienne Leonard for convincing me to learn to row, for quickly realising my potential in the sport, for convincing me to run for women's captain in Clare Hall Boat Club, and for all her advice after I was elected. Thanks also to my First Ladies Crew in Clare Hall Boat Club, who won blades with me in May Bumps 2009 and 2010. Thanks also go to Dr. Elisa Dassie, without whom Clare Hall rowing would not have been the same. Thank you, Elisa and Adrienne, not only for all the good times on the water, but also at the May Balls. I would further like to thank Nick Acock, who answered all my questions about trialling for the University, which leads me to my time at Cambridge University Women's Boat Club. I would like to thank Sarah Allen for running the squad in 2009/10 and for convincing me that I have what it takes. Most importantly, I would like to thank my lovely 2010 Blondie crew, especially Dr. Stacey Hickson and Claudia Cattachio, who made this time of extreme training, working and sleeping schedules one of the best times of my life. Finally, I would like to thank Stacey, Rebecca, Sophie and Mat for the hard training in the IV+ and for winning the gold medal at the British University and College Championships Regatta with me!

Last, but by no means least, I would like to acknowledge the people who supported me in my personal life. First I would like to thank my parents who have made it possible for me to study here and especially to my mum who has always supported me in my professional choices, even if it meant that I'd be living in another country. Finally, I would like to thank Tjaart for his continuous support, especially in these last months before my submission.

Abstract

Enzymes are the cell's executive molecules that catalyse the chemical reactions required for an organism to function. Some enzymes require cofactors (metal ions or organic small molecules) for catalysis. Approximately half of all enzyme reactions crucially depend on organic enzyme cofactors. In order to better understand the chemistry of life, it is important to understand the properties, evolutionary context and functional roles of organic enzyme cofactors.

The aim of this work is to investigate this important group of compounds. Which molecules are organic enzyme cofactors? What is their molecular function and enzymatic mechanism? What are their physicochemical properties? How are they different from other metabolites in the cell? Are there intrinsic groupings among them? Which enzyme reactions use them? How variable is their conformation in crystallographic protein structures and in solution? How does their conformation vary among homologous proteins? These questions have been investigated using computational methods.

The data were extracted manually from the scientific literature and automatically from web resources and has been stored in a database. Statistical methods including uni- and multivariate data analysis have been applied to analyse one-dimensional descriptors, two-dimensional structures and functional roles of the cofactor molecules, while the conformational analysis further uses three-dimensional superposition.

The results show that organic enzyme cofactors are slightly larger and more polar than the other metabolites although their physicochemical properties sample all of metabolite space. Intrinsic groupings can be identified based on descriptors, two-dimensional structures and

molecular functions. Some organic cofactors share common building blocks, which complete and partially complement the functional profile of catalytic amino acids and metal ions. The conformational variability of a cofactor depends on its group membership and can vary substantially from its conformational variability in solution.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

ATP	Adenosine triphosphate
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
GTP	Guanosine triphosphate
HET	Hetero atom three-letter-codes from the PDB format
HIV	Human immunodeficiency virus
MB	Megabytes
mRNA	messenger ribonucleic acid
NCBI	National Center for Biotechnology Information(USA)
PC(A)	Principal Component (Analysis)
RNA	Ribonucleic acid
ROS	Reactive oxygen species

Cofactor abbreviations and HET codes

AdoMet	S-adenosylmethionine
ALAS-1	Non-specific 5-aminolevulinate synthase

ASC	Ascorbic acid (vitamin C)
B12	Adenosylcobalamin (Vitamin B_{12})
BH4	Tetrahydrobiopterin
BTN	Biotin
COA or CoA	Coenzyme A
CoB	Coenzyme B
COM	Coenzyme M
CoM	Coenzyme M
CoQ	Coenzyme Q, Uniquinone
CTQ	Cysteine tryptophylquinone
DPM	Dipyrromethane
F43	Factor F430
F430	Factor F430
FAD	Flavin adenine dinucleotide
FMN	Flavin mono-nucleotide
GASSAG	Glutathione amide disulfide
GSH	Glutathione
GSSG	Glutathione disulfide
GTT	Glutathione
HEA	Heme A
HEM	Heme
HEO	Heme-O

LA	Lipoic acid
LPA	Lipoic acid
LTQ	Lysine tyrosylquinone
MIO	4-methylideneimidazole-5-one
MoCo	Molybdenum cofactor
MQ	Menaquinone (Vitamin K)
MQ7	Menaquinone-7
MTE	Molybdopterin
NAD	Nicotinaminde-adenine dinucleotide phosphate
NAD	Nicotinaminde-adenine dinucleotide
NAD(P)	Nicotinaminde-adenine dinucleotide (phosphate)
OQs	Orthoquinone residues
PLP	Pyridoxal 5'-phosphate
PNS	Phosphopantetheine
PQQ	Pyrroloquinoline quinone
SAM	S-adenosylmethionine
SRM	Siroheme
ThDP	Thiamine diphosphate
THF	Tetrahydrofolic acid
TP7	Coenzyme B
TPQ	Topaquinone
TTQ	Tryptophan tryptophylquinone

U10	Ubiquinone-10

Databases and web resources

CATH	Classification of protein domain structures (Class, Architecture, Topology, Homologous superfamily)
ChEBI	Chemical Entities of Biological Interest
E.C.	Enzyme Commission: hierarchically classifies enzymes based on their overall reaction
IntEnz	Integrated relational Enzyme database
KEGG	Kyoto Encyclopedia of Genes and Genomes
MACiE	Mechanism, Annotation and Classification in Enzymes
PDB	Protein Data Bank
PDBe	Protein Data Bank in Europe
PDBeChem	Dictionary of chemical components (ligands, small molecules and monomers) referred in PDB entries and maintained by the ww-PDB
PDBsum	At-a-glance overview of every macromolecular structure deposited in the Protein Data Bank (PDB), giving schematic diagrams of the molecules in each structure and of the interactions between them
PROCOGNATE	Database of cognate ligands for the domains of enzyme structures
wwPDB	world wide Protein Data Bank

Vitamins ..

Vitamin B_1	Thiamine diphosphate is biosynthesised from vitamin B_1

Chapter 1

Introduction

1.1 Motivation

Enzyme reactions have been widely investigated for many decades, yet their modes of action, mechanisms and three-dimensional structures are still poorly understood in many cases. New insights are therefore still published regularly. Enzymes are the main effector molecules in the cell and without them cellular metabolism, as we know it, would not be sustainable. An estimated 45% of all known enzyme reactions (as defined by their Enzyme Commission number, NC-IUBMB & Webb, 1992) rely on at least one cofactor to perform their task.

Cofactors are molecules that actively assist the catalytic amino acids in enzyme catalysis. Generally there are metal ions and small organic molecules performing this cofactor function in enzymes. In this thesis, organic enzyme cofactors will be investigated as one of the three catalytic entities (catalytic amino acids, metal ions and organic cofactors) that enable enzymatic reactions, and thus the chemistry of life.

Although a lot of research is focused on single organic cofactor molecules or groups of them, to the author's best knowledge no systematic study of organic enzyme cofactors from a general perspective has been published to date[1]. Understanding how enzymes catalyse the vast majority of the chemical reactions required for life, with astonishing efficiency, selectivity of substrates and under

[1] except for the author's own publications

physiological conditions, is not only fascinating but also of great interest for industrial applications. For example, enzymes can be used to selectively synthesise a specific stereoisomer of a desired chemical compound. Further, enzymes themselves are used as an additive to detergents like washing powder, or in food processing, for instance in the production of diary products. In pharmaceutical research, too, enzymes are of great importance, as the majority of all approved drug targets are enzymes (Overington *et al.*, 2006). The incentive to gain a better understanding of enzyme catalysis is strong from a scientific as well as an industrial perspective, and the motivation to learn about organic enzyme cofactors is rooted in their large contribution to Nature's catalytic toolkit.

1.2 Proteins and enzymes

Enzymes are, like all proteins, made up of amino acids connected through peptide bonds (Bugg, 2004, 8). There are 20[1] standard amino acid building blocks.

1.2.1 Introduction to proteins

When a protein is biosynthesised from the mRNA template in the ribosome, specific amino acids enter the active site of the ribosome based on the code on the mRNA template. Neighbouring amino acids form peptide bonds and a peptide chain (*primary structure* of the protein) is produced (Patthy, 1999, 8-9,12). This one-dimensional chain locally adopts one of three general conformational states in folded[2] proteins, which is referred to as the protein's *secondary structure*: a right-handed helix (α-helix), sheet-like arrangements of sections of the peptide chain (β-sheet) or β-turns (Patthy, 1999, 22-24). Due to the different physicochemical properties of the amino acids in the chain, the three-dimensional folding of a protein chain (*tertiary structure*) is mainly driven by entropy and the hydrophobic effect (Fersht, 1999, 508-515). The hydrophobic amino acids are situated mostly in the core of the folded protein, whereas the polar and charged amino acid

[1] The vast majority of proteins consists of the 20 standard amino acids. However, some organisims use other amino acids, such as selenocysteine and pyrroyllysine.

[2] Not all residues are in these three well-structured states. About 30% are unstructured, typically in extended loop regions.

residues are more likely to be found on the surface (Branden & Tooze, 1991, 38). Both the hydrophobic effect and the change in entropy are based on the high water content of the cell: the hydrophobic residues maximise their favourable interaction with each other by minimising their contact surface with polar water (hydrophobic effect), and the order among the water molecules is lower when the protein is folded, thus resulting in a loss of entropy (Fersht, 1999, 508-510) for the water molecules.

A folded unit of protein in three-dimensional space is usually referred to as a protein domain (Branden & Tooze, 1991, 26). Note that there is no one-to-one relationship between protein chains and domains: one protein chain may encode two or more domains (Branden & Tooze, 1991, 26). A functional protein may further require several domains (Patthy, 1999, 142-143), and in fact several amino acid chains to work together. If a protein's chains adopt a specific orientation towards each other, this is referred to as the protein's quaternary structure (Fersht, 1999, 24).

Several databases aim to classify protein domains, for instance SCOP (Structural Classification Of Proteins, Andreeva *et al.*, 2008), CATH (Greene *et al.*, 2007) and Pfam (Protein families, Finn *et al.*,2010). In this work, the CATH classification of protein structures is used to identify domains in protein structures. CATH hierarchically classifies protein structural domains. The first four levels of the hierarchy are <u>C</u>lass, <u>A</u>rchitecture, <u>T</u>opology and <u>H</u>omologous superfamily. Hence, the CATH definition of structural domains, as stated on their web page (CATH, 2009), is used here. The CATH database uses a computer-assisted but manually curated approach to determine domains in protein structures.

It is worth noting that proteins are not static, or fixed to one exact conformation in three-dimensional space. While globular proteins are often more densely packed and rigid in the core, the solvent-exposed fragments often vary considerably in conformation (Fersht, 1999, 24). Although the fold of a protein is generally defined by its primary structure, proteins have a certain degree of flexibility. Conformational changes like inter-domain or major loop flexibility may occur upon ligand binding, post-translational modification or during catalysis and thus often have functional relevance (Fersht, 1999, 369).

1.2.2 Introduction to enzymes

Enzymes are all those proteins that catalyse chemical reactions, *i.e.* transformations from one or more substrates to one or more products (Bugg, 2004, 1,8). The active site of the enzyme is the site where the substrates bind to the enzyme and the transformation occurs. Mostly, the active site is located in one of the largest clefts in the enzyme's surface and often it is deeply buried (Bartlett *et al.*, 2002a).

Enzymes are catalysts, and thus they accelerate chemical reactions. There are many different hypotheses that aim to explain how enzymes work (Silverman, 2004, 174). The most commonly used and accessible hypothesis is transition state theory, in which enzymes work by lowering the activation energy barrier (Bugg, 2004, 31). The activation energy is "the difference in free energy between the reagent(s) and the transition state for the reaction" (Bugg, 2004, 31). Thus, the activation energy is a high energy state of the reaction intermediate that the reaction coordinate passes through on the way from substrates to products. The activation energy can be lowered by stabilisation of the transition state or by using a lower energy reaction pathway (Bugg, 2004, 31). These goals may be achieved by several means. The shape of the active site (binding pocket, as defined by Kahraman *et al.*, 2007) may restrict the conformations that the substrates can adopt. Further, the electrostatic environment and the pH value of the active site may be adjusted depending on which amino acids line the surface of the active site cavity. Generally speaking, the purpose of the active site is to create a favourable environment that makes the chemical reaction catalysed by the enzyme more likely to happen.

Enzymes are classified by the Enzyme Commission (E.C., NC-IUBMB & Webb, 1992) based on the overall chemical transformation they catalyse. The E.C. number is a four level hierarchical classification of an enzyme's overall reaction. The E.C. class (level 1) denotes the general type of reaction, of which there are six (1: oxidoreductases, 2: transferases, 3: hydrolases, 4: lyases, 5: isomerases, 6: ligases). The next two classification levels (level 2: sub-class and level 3: sub-sub-class) depend on the class, but generally describe the chemistry of the overall reaction in ever-greater detail. The fourth level gives a serial number for each enzyme reaction, effectively describing the substrate specificity of the enzyme. It is crucial to note that one E.C. number does not denote an enzyme, but an overall chemical reaction. Thus, several enzymes, which may be non-homologous, may be identified by the same E.C. number if they catalyse the same overall reaction (Galperin *et al.*, 1998). To what extent this happens has been investigated in our research group and others (Galperin *et al.*, 1998; Holliday *et al.*, 2011). Holliday *et al.* (2011) have estimated the number of different enzyme families per E.C. node to be between 1.93 and 3.45.

1.3 Cofactors

Cofactors are molecules or ions which are essential for some enzymes. For instance, some vitamins have been identified early in the 20th century as cofactors or coenzymes (Bugg, 2004, 3). As the scope and exact terminology of the terms cofactor and coenzyme varies, the first task was to establish a working definition of what exactly was meant by these terms.

1.3.1 Cofactor definition for this thesis

In mid 20th century scientific publications, the word cofactor denoted any molecule that is required for an enzyme reaction to occur (often *in vitro*, e.g. Silverman & Nandi, 1988) apart from the substrate(s) and the (apo-) protein consisting only of the 20 amino acids. With increasing structural knowledge, the term "cofactor" was refined to small molecules that are necessary for catalysis and present in the active site. This already excludes regulatory site binders, although these had often been referred to as cofactors in publications (e.g. Larsson & Reichard, 1966). Often, the terms "coenzyme", "cofactor" and "cosubstrate" are used interchangeably, which has, in some circles, lead to the assumption that coenzymes are catalysts in the pure chemical sense and thus have to emerge unchanged after each catalytic cycle, much like an enzyme does. To define what a cofactor should be in this study, the IUPAC Gold Book (McNaught & Wilkinson, 1997; Union of

Pure & Applied Chemistry, 2005–2009) was consulted, which states that cofactors are "organic molecules ... or ions ... that are required by an enzyme [for] its activity." This definition also states that a "cofactor binds with its associated protein (apoenzymes), which is functionally inactive, to form the active enzyme (holoenzyme)" (Union of Pure & Applied Chemistry, 2005–2009).

However, this definition is further refined for this work, by requiring that a cofactor must be present at the catalytic site. This is necessary, because cases of constitutive allosteric regulation exist (e.g. phosphobilinogen synthase, E.C. 4.2.1.24, uses magnesium ions as allosteric activators), but they are not considered to be cofactors, as they do not play an active role in the catalytic process, *i.e.* they are chemically inert.

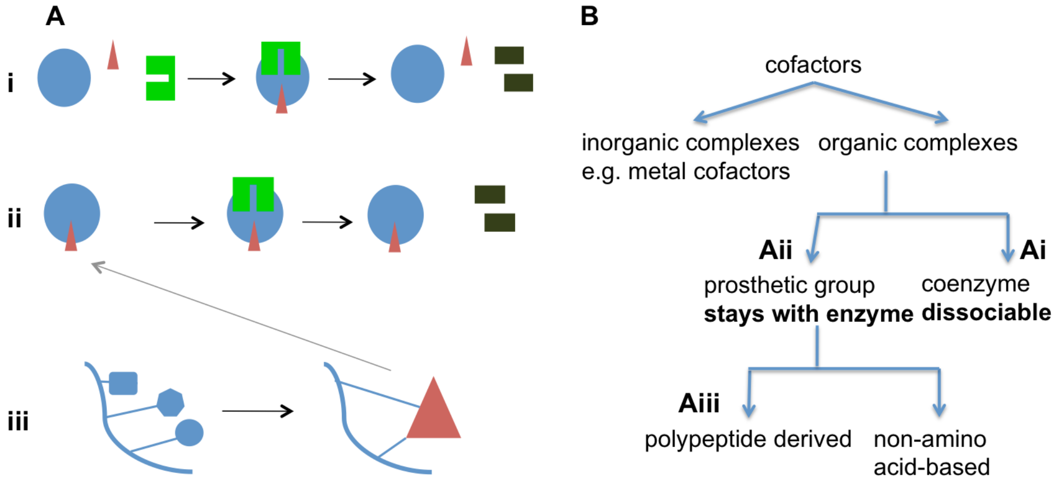


Figure 1.1: A: Schematic of the different types of cofactor-roles, i: coenzyme role, ii: prosthetic grup role, and iii: polypeptide-derived prosthetic group. Blue circle: enzyme, red triangle: cofactor. The substrate is shown in green and the product in gray. B: Classification of cofactors.

To further differentiate cofactors, it is worth following up on another sentence in the Gold Book definition of cofactors: "They may be attached either loosely or tightly (prosthetic group) to the enzyme" (Union of Pure & Applied Chemistry, 2005–2009). Therefore, two kinds of cofactors are distinguished: coenzymes and prosthetic groups. The Gold Book defines a coenzyme as "the dissociable, low-relative-molecular-mass active group of an enzyme which transfers chemical groups, hydrogen or electrons" (Union of Pure & Applied Chemistry, 2005–2009). In contrast to this dissociable form of a cofactor, a prosthetic group is "the non-amino acid portion of a conjugated protein" (Union of Pure & Applied Chemistry, 2005–2009). Although there are also prosthetic groups that are not cofactors (e.g. retinal in light receptors), only those prosthetic groups that are located in the active site of an enzyme (and not in other proteins) are denoted cofactors. As mentioned already in the Gold Book definition of the term cofactor, a prosthetic group "may be attached either loosely or tightly ... to the enzyme" (Union of Pure & Applied Chemistry, 2005–2009). For our purpose, a prosthetic group is therefore distinguished from a coenzyme in that – once inserted into the apoenzyme – it stays with the enzyme over many catalytic cycles, possibly until the enzyme is degraded. The coenzyme, on the other hand, binds to the enzyme at the beginning of each catalytic cycle and leaves at the end of it. It may therefore be alternatively referred to as cosubstrate, although this term is not defined in the Gold Book.

While the Gold Book states that prosthetic groups are non-amino acid portions, here cofactors that are (partially) made of amino acids are not excluded from the definition. Coenzyme A or the structurally related molecule pantothenic acid, for instance, are partly composed of standard amino acids, but build a new molecule that is not assembled by the ribosome. These molecules are included in the definition of cofactors for this work. Further a special kind of prosthetic groups are included, which are the polypeptide-derived cofactors. They arise from post-translational modifications of standard amino acids (e.g. TPQ, Schwartz & Klinman 2001) or from auto-catalysis of two or more standard amino acids in the apoenzyme (e.g. MIO, Christianson *et al.*, 2007b). This does not include modifications for signalling purposes (e.g. phosphorylation, acetylation etc.) or selenocysteine. The hierarchy of the terms cofactor, coenzyme, prosthetic group and polypeptide-derived cofactor is summarised in figure 1.1.

1.3.2 Ambiguous cases

In some cases it is difficult to decide if a compound is a coenzyme or a substrate to the reaction. Examples are molecules like NAD(P)H or ATP. In this work, NAD(P)H is included as a cofactor, because it fulfils all the criteria (transfers electrons, is dissociable, is not a standard amino acid and is actively involved in catalysis) and moreover because this function is the molecule's main purpose in the cell. In contrast, ATP does not – in the vast majority of cases – fulfil all the criteria: although it is dissociable, not a standard amino acid and actively involved in catalysis, it usually does not *transfer* the phosphate group, but instead it has one of its P-O bonds cleaved and ADP plus inorganic phosphate (or AMP plus diphosphate) released in order to provide energy for otherwise endergonic reactions. The molecule also functions as an allosteric regulator of some enzymes (Datta, 1970; Larsson & Reichard, 1966). Since ATP's primary purpose in the cell is to provide energy and to be a building block for RNA (and DNA) sequences, it is not considered a cofactor in this work.

Some molecules may act as cofactors, but have additional roles, for instance in small molecule transport. Heme, for example, famously transports oxygen within the blood stream, but does not catalyse a metabolic reaction in the corresponding protein hemoglobin. In other reactions, however, heme molecules are actively involved in a catalytic mechanism (for instance in the peroxidase reaction, E.C. 1.11.1.7).

There are several example reactions of molecules that can be easily mistaken for cofactors although they are not, according to the above definition. A classic example is coenzyme A: its main purpose in the cell is to transfer acetyl-residues (C_1) to a substrate, but it also serves as an acyl-carrier (longer C_n chains). In the latter function, coenzyme A does not act as a coenzyme, because the sulfur atom, which executes the group-transfer, is not involved in the reaction. An example for this is acetyl-CoA in the mechanism of acetyl-CoA C-acyltransferase (E.C. 2.3.1.16, see MACiE (Holliday *et al.*, 2005) entry M0077 for mechanism). In that reaction (acyl-CoA + acetyl-CoA → CoA + 3-oxoacyl-CoA), it is the other substrate, acyl-CoA, that transfers a group to cysteine residue, which eventually transfers it to the group acceptor acetyl-CoA. Therefore, acyl-CoA is considered to be a cofactor in this reaction, but acetyl-CoA is not. Another interesting example is pyrogallol hydroxytransferase (E.C. 1.97.1.2). Here, the same molecule appears on the substrate and on the product side (A+B → B+C), as shown in figure 1.2. One might argue that B is therefore a cofactor, because the molecule emerges from the reaction unchanged, like a real catalyst. However, after taking a closer look at the mechanism, it becomes clear that molecule B is in fact a substrate and not a cofactor, because the majority of the atoms in B (green in figure 1.2) on the left-hand side end up as atoms of C on the right-hand side. If B was a cofactor, it should be composed of the identical atoms on both sides of the reaction arrow (with the exception of protons, electrons and the groups that it might transfer).

Considering all of the above, table 1.1 contains all the molecules that are considered as cofactors in this work.

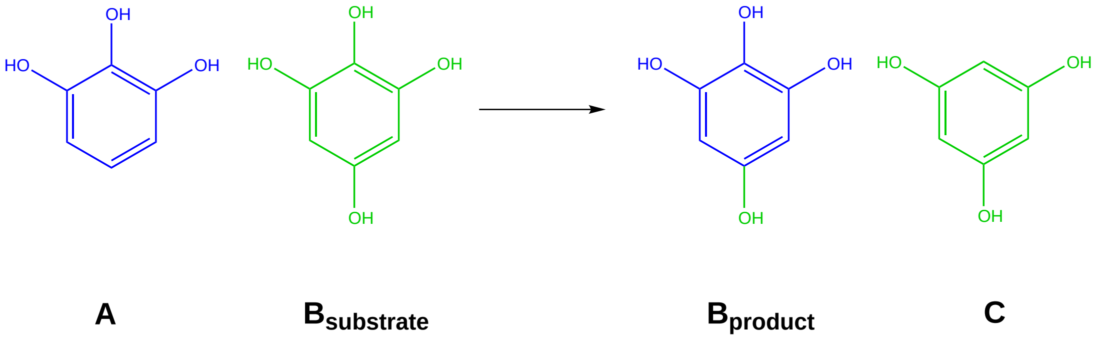
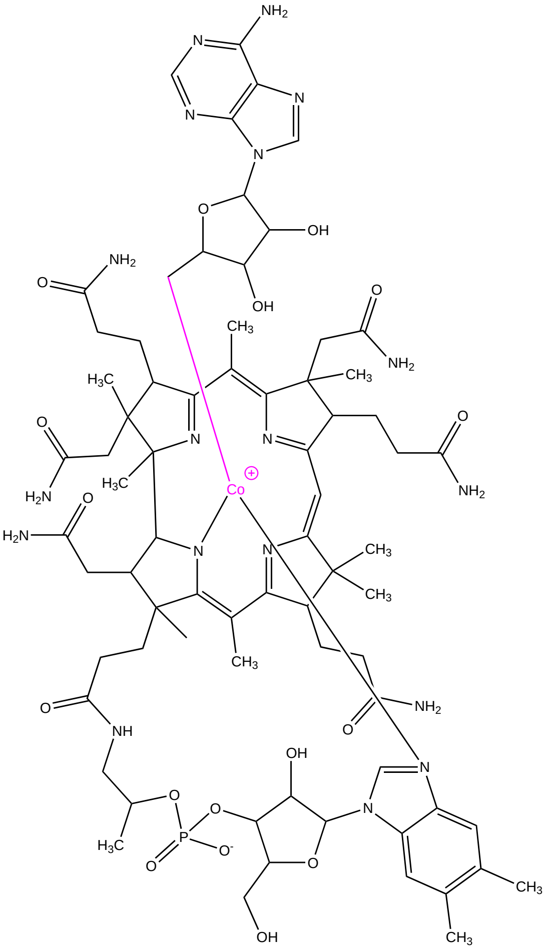
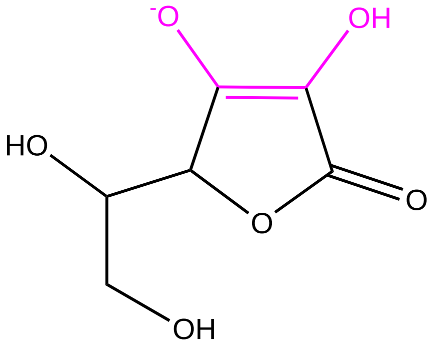
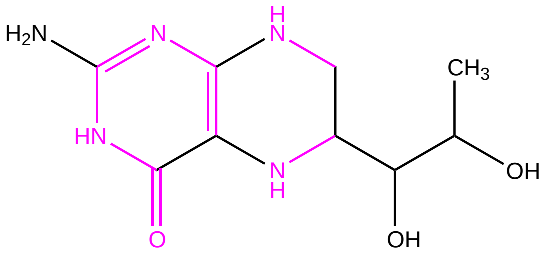
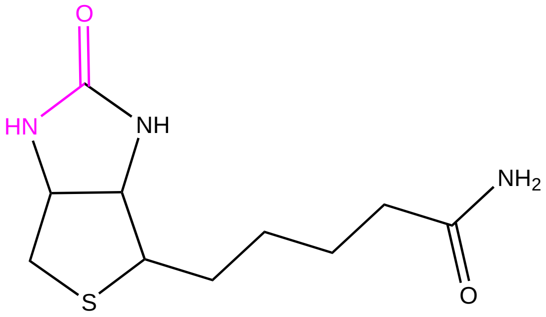
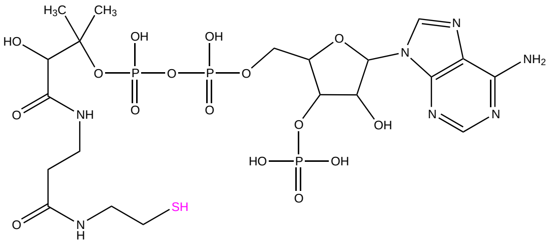
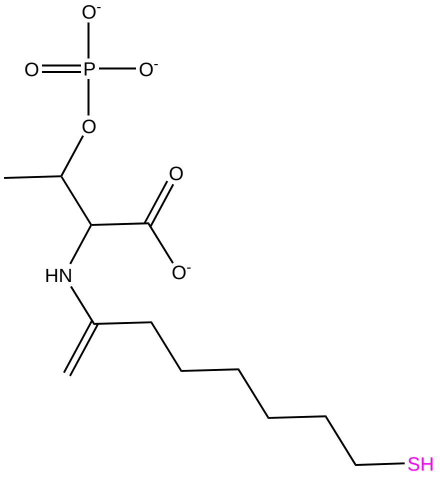


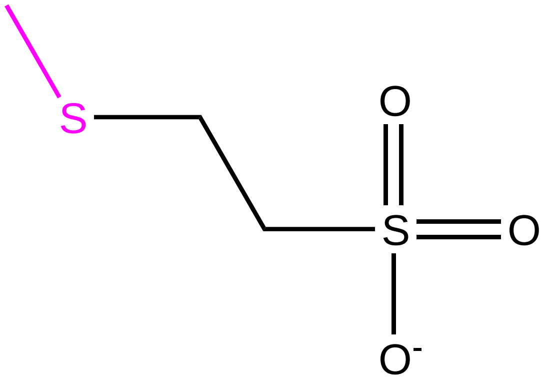
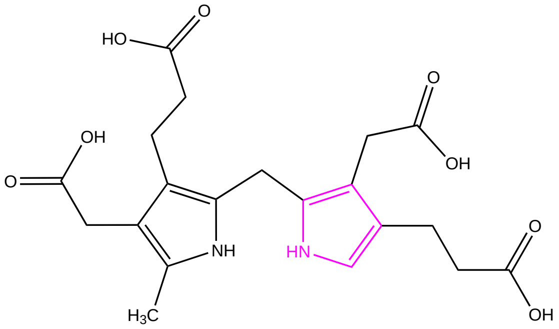
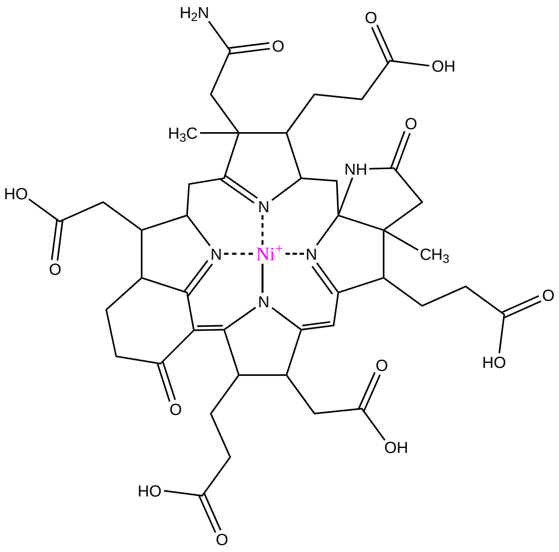
Figure 1.2: Overall reaction of pyrogallol hydroxytransferase (E.C. 1.97.1.2, MACiE entry M0146. Identical heavy atoms on substrate and product sides are coloured with identical colours.

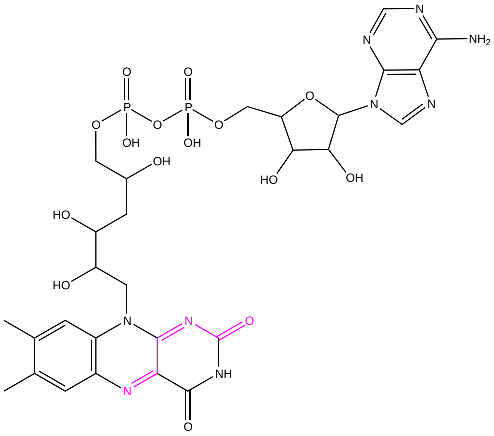
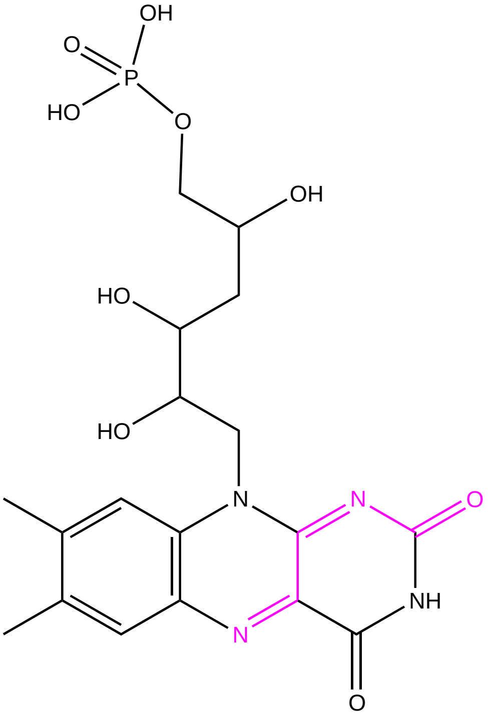
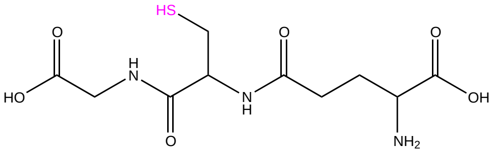
Therefore, to summarise, our working definition of an organic cofactor is an organic molecule, which is located at the active site of an enzyme and is actively involved in biocatalysis. Cofactors are required by an enzyme for its activity and standard amino acids are excluded from the definition. A cofactor's main task in the cell is to assist in the catalysis of enzyme reactions. It can be either dissociable (coenzyme) or attached tightly (prosthetic group) to the enzyme. A cofactor transfers small chemical groups, electrons or protons and may or may not emerge unchanged from the reaction.

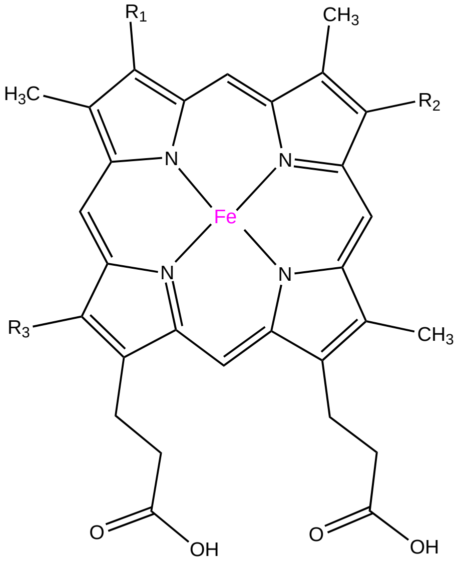
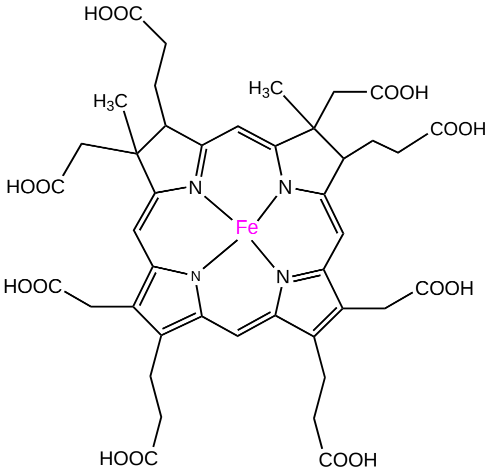
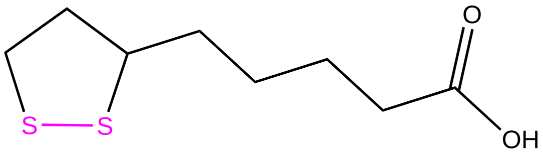
Individual cofactors have long been identified and extensive experimental studies have been performed, investigating their involvement in nutrition and disease. The objective of this work is to understand the use and distribution of cofactors in enzymes, and to explore their physicochemical properties, chemical structures, structural conformations as well as their function. Here for the first time, all the organic enzyme cofactors are brought together. All this information is collected and presented with key facts about each cofactor from primary literature and stored in a publicly available database that can be viewed online at http://www.ebi.ac.uk/thornton-srv/databases/CoFactor.

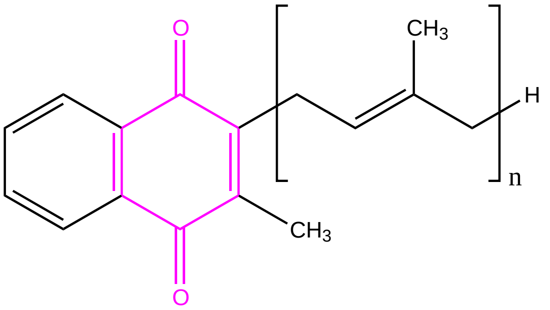
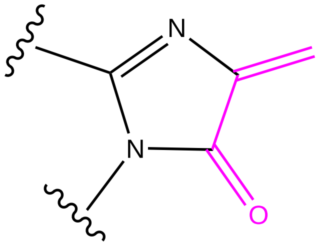
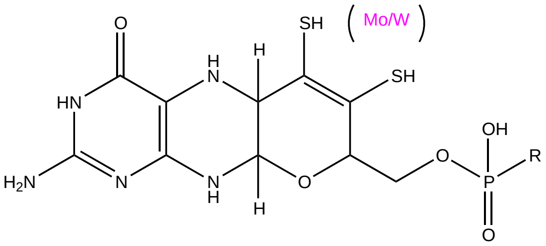
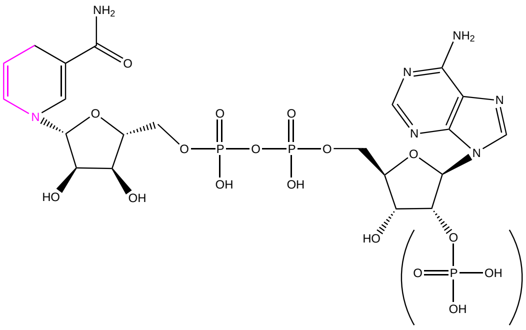
CoFactor name	Function	HET
Adenosylcobalamin 	Rearrangements Bond cleavage Group transfer (-CH_3)	B12,COB
Ascorbic acid 	Redox	ASC

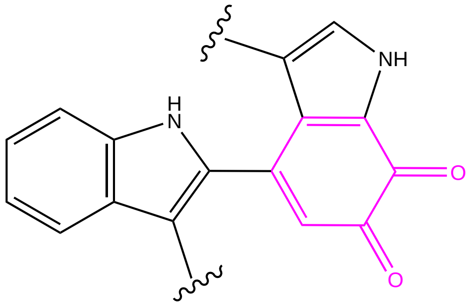
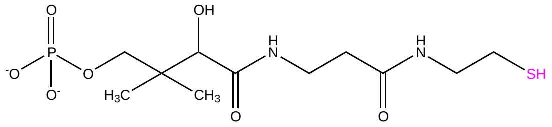
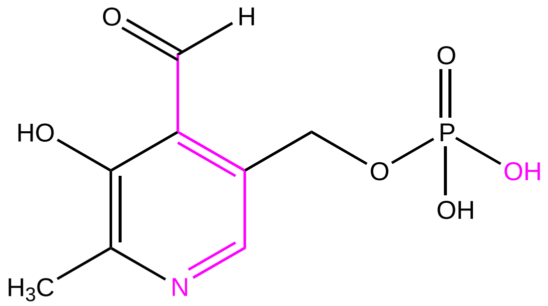
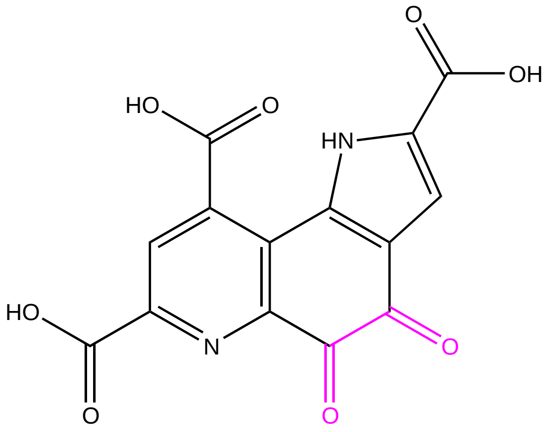
Biopterin 	Group transfer (-OH) Redox	H4B,BH4, THB
Biotin 	Group transfer (-CO_2)	BTN
Coenzyme A  H3C CH3 HO OH OH O N N NH2 O P O P O O O O NH N N OH O HO P OH O O N H SH	Group transfer (-CR)	COA,COZ
Coenzyme B 	Redox	TP7

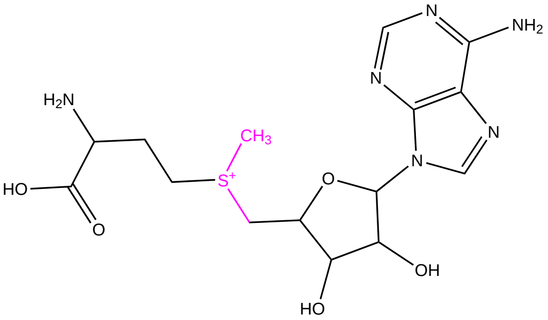
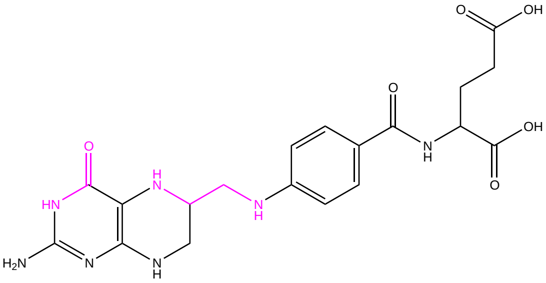
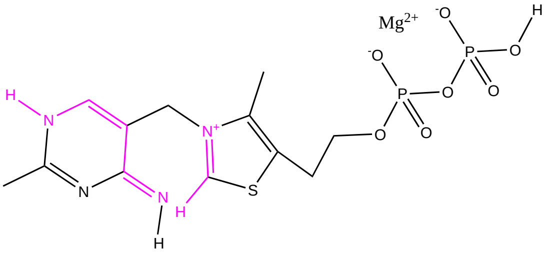
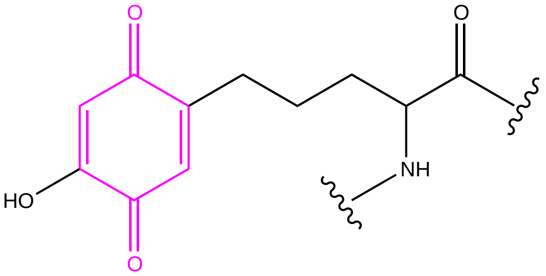
Coenzyme M 	Group transfer (-CH_3)	COM
Dipyrromethane 	Polymerisation	DPM
Factor F430 	Redox	F43

Flavin adenine dinucleotide  	Redox Activation	FAD,FDA
Flavin Mononucleotide  	Redox	FMN
Glutathione  	Redox	GSH,GTT

Heme  R1 CH3 H3C R2 N N Fe N N R3 CH3 O OH O OH	Redox	HEB,HEM, HEA,HEC, HEO
Siroheme  HOOC H3C H3C COOH COOH HOOC N N Fe HOOC N N COOH HOOC COOH	Redox	SRM
Lipoic acid 	Redox Group transfer ($C(CH_3)O$) Mobility	LPA

Menaquinone  	Redox	MQ7
MIO  	Group transfer (NH_2)	MDO
Molybdopterin  O SH Mo/W H H N SH HN OH R H2N N N H O O P H O	Redox	MGD,MTE, PGD,PTT, 2MD,MSS, PCD,PTE
Nicotinamide-adenine dinucleotide  NH2 O N O HO OH O P O P O OH OH NH2 N N N N O HO O O P OH OH	Redox	NAH,NAP, NAD,NAI, NDP

Orthoquinone residues (LTQ, TTQ, CTQ) 	Redox	–
Phosphopantetheine 	Group transfer (RC=O), Mobility	PNS
Pyridoxal 5'-phosphate 	Group transfer (NH$_2$ or CO$_2$), Bond cleavage/formation	PLP,PMP
Pyrroloquinoline Quinone 	Redox	PQQ

S-adenosylmethionine 	Redox Group transfer(CH_3)	SAM
Tetrahydrofolic acid 	Group transfer (CH_3) Activation	THF,MHF
Thiamine diphosphate 	Bond cleavage	TDP,TPP
Topaquinone 	Redox	TPQ

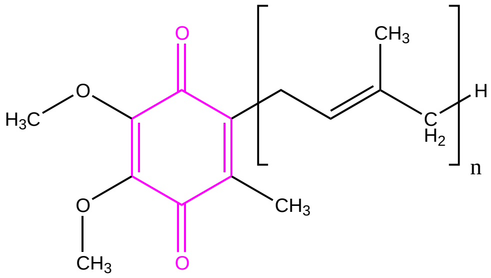
Ubiquinone (Coenzyme Q)  O CH3 H3C O H C H2 n O CH3 CH3 O	Redox	U10,UQ1, UQ2,UQ5, UQ7

Table 1.1: Overview of all organic cofactors considered in this work: their names, two-dimensional structures (active atoms highlighted in magenta), their functions and PDB HET codes.

1.4 Existing work

The amount of published research on each single cofactor is overwhelming and often spans decades. Thus summarising the entire literature on all cofactors exceeds the scope of this thesis. The CoFactor database (see chapter 3) and a summary of the hand-curated information stored there (see section 3.2) provides a rough overview for each cofactor and links to the scientific literature are provided there.

1.4.1 Biocatalysis, enzyme design and evolution of cofactors

In this section, an overview of the basics as well as the current topics in enzyme research is given. The aim is to demonstrate the applictions of this research field, and to show its importance and possible impact on society in the future.

1.4.1.1 Biocatalysis

Catalysis of chemical reactions in biological systems has been found to be performed by catalytic RNA molecules or enzymes. There are various theories that explain how enzymes generally transform substrates to products. From a kinetic point of view, the concentration of the substrates and products and their respective free energy defines the equilibrium for the reaction, while the activation energy and thus the transition state define the rate of the reaction. From thermodynamics it is known that the change in free energy (ΔG) depends on the temperature T, the change in entropy (ΔS) and the change in enthalpy (ΔH), which occur over the course of the reaction. Equation 1.1 describes the relationships between these variables (Fersht, 1999, 55-56).

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

The most popular theory to date of how enzymes generally work is the transition state theory, which has been summarised in section 1.2.2. Later research suggests that proton tunnelling is an alternative mechanism with which the enzyme may achieve the transformation by passing through the activation energy barrier, rather than over it (Benkovic & Hammes-Schiffer, 2003; Klinman, 2009; Sutcliffe & Scrutton, 2000). Quantum tunnelling has been shown to occur for small particles with very little mass like electrons and it is an established mechanism for biological electron transfer (Markus & Sutin, 1985). The recent research suggests that this might also happen for larger particles like protons or hydrogen atoms, though over smaller distances due to their larger mass. In combination with the changing height of the activation energy barrier due to the protein's conformational flexibility, hydrogen or proton tunnelling is considered to be a possible mechanism by some scientists (Sutcliffe & Scrutton, 2000) for at least a subset of enzymes, although opinions in the field diverge (Kamerlin *et al.*, 2010).

1.4.1.2 Recent progress in biocatalysis and enzyme design

In recent decades, the idea to engineer and modify naturally occurring enzymes has inspired researchers from various backgrounds. Potential applications of artificially modified enzymes that are slightly adapted and tailored to a specific task promise advances in stereospecific synthesis of drugs for medical applications, chemicals for industrial applications, or even biofuel cells (Kim *et al.*, 2006).

The relatively mild conditions (compared to some chemical synthesis processes), in which enzymes catalyse reactions, and the high catalytic turnover are two of the desirable properties of enzyme catalysis (Cirino & Sun, 2008). As biosynthetic pathways are modular (one enzyme as one module), the idea to combine modules from biologically separated pathways to form new natural product-like molecules has been discussed (see review by Zhou *et al.*, 2008). Realising this idea may be difficult, for instance if the native substrate of the template enzyme is slightly different from the molecule that needs to be transformed in the industrial application process. Directed evolution has been used to address this problem in order to adapt an enzyme's substrate specificity. This technique may further be applied to improve other catalytic properties of an enzyme's ability to perform the target reaction (Rubin-Pitel & Zhao, 2006). Coward-Kelly & Chen (2008) have recently reviewed tools and technologies that aim to solve the current problems in biocatalysis. They conclude that techniques from areas such as nanotechnology, bioinformatics and protein engineering are becoming more popular and yield satisfactory results. This is facilitated by recent advances in engineering platforms, which have recently been reviewed by Cirino & Sun (2008).

Most recently, miniaturised reactors have been used to parallelise enzyme reactions and thus increase the throughput. A more detailed review of the advances of miniaturisation in biocatalysis was published by Fernandes (2010).

In some areas, enzyme design has already been implemented. Proteases have been used as biocatalysts in laundry detergent since the 1960s, to facilitate the removal of food stains (Banik & Prakash, 2004). To achieve this, the enzyme's stability needs to be optimised for higher pH values, and when washing at higher temperatures, the thermostability is adjusted as well. This is necessary for the enzyme to work under the conditions in a washing machine, rather than in a living cell.

In pharmaceutical research, enzymes such as the non-ribosomal peptide synthases have been shown to produce bioactive natural products including some antibiotics. These enzymes have been recognised as a modular system of "sub-enzymes" that may potentially be engineered to produce new natural product-like compounds. A full review of non-ribosomal peptide synthases was published by Sieber & Marahiel (2005).

Enzyme-based biofuel cells convert chemical energy to electrical energy. They oxidise for example sugars instead of fossil fuels, which is an important advantage over burning fossil fuels with respect to counteracting the effects of human-induced climate change. They have been heavily investigated, since the system offers renewable catalysts that operate at room temperature (Minteer *et al.*, 2007). In this context mediators are those compounds that shuttle electrons, which are often organometallic complexes, much like some organic cofactors are.

Some cofactors, like factor F430, coenzyme B and coenzyme M, take part in a methane-generating reaction (coenzyme-B sulfoethylthiotransferase, E.C. 2.8.4.1) in methanogenic bacteria. This reaction is also important in considering renewable energy sources and climate change (Dey *et al.*, 2010). All these ideas connect enzyme design for industrial applications with the research on organic enzyme cofactors.

Although the technical details of how enzymes may be artificially modified exceed the scope of this thesis, the idea of enzyme design is relevant for this work. Organic cofactors have already played important roles in enzyme design experiments, for instance when an enzyme was modified to use NADP instead of NAD (Nair & Zhao, 2007). More crucial changes in cofactor usage might be an interesting tool in future enzyme design projects.

1.4.1.3 Evolution of cofactors

One hypothesis of how early life might have evolved on Earth is the RNA-world hypothesis (Muller, 2006). According to this proposal, RNA molecules once functioned as molecules for storing genetic information as well as for executing catalysis. These tasks are thought to be mainly performed by DNA and proteins, respectively, in cells today. The central position of RNA molecules in the flow of information in today's living cells has often been seen as supporting evidence for the RNA-world hypothesis. Further, many cofactors have RNA-portions: from whole RNA nucleotides like NAD or FAD, to parts of nucleotides like S-adenysolmethionine and vitamin B_{12}. This may be considered as evidence in today's cellular chemistry for remnants from the RNA-world, where these cofactors may have been recruited by RNA molecules to perform catalysis. These RNA-portions (even phosphate groups as found in thiamine diphosphate) may have been ways to connect a chemically and catalytically useful compound to the ribozyme. It is important to note that the RNA-world hypothesis, although very appealing, is not (yet) generally accepted and may change with new scientific evidence. A comprehensive review of the theories on the emergence of life on earth has been published by Lahav (2001).

The use of organic cofactors by catalytic RNA molecules in a hypothetical RNA-world was investigated by Jadhav & Yarus (2002) and found to be likely, at least for the nucleotidyl coenzymes. Although this aspect of organic cofactors is a fascinating subject, the focus of this thesis is mainly on organic cofactors found in enzymes that exist to date.

1.4.2 Existing data resources

The rise of the Internet has enabled researchers to provide and retrieve data about enzymes and the compounds involved in their reactions from databases with online access. This development is of great use to both the experimentalists, who use them to learn details about the enzyme(s) that are of importance to their research question, and to the research informaticians, who might want to investigate one aspect of enzymes or their compounds in a large data set. The next two sections thus aim to provide an overview of the available data resources.

1.4.2.1 Compound resources

Over the last decade, major efforts have been undertaken by the large research institutes all over the world to make the chemical knowledge more accessible to large-scale analysis. In Japan, the KEGG project offers a wide range of drug, enzyme and compound related databases (Kanehisa & Goto, 2000). KEGG COMPOUND is the most relevant of these databases to this work, as it holds information about all the compounds that are substrates and products in KEGG ENZYME. The latter is an enzyme reaction database that follows the E.C. classification (NC-IUBMB & Webb, 1992) with added integrated information, for example about links to pathways or drug molecules. In the United States, the National Center for Biotechnology Information (NCBI) hosts the PubChem (Bolton *et al.*, 2008) databases and in Europe, the European Bioinformatics Institute (EBI) runs ChEBI (Degtyarenko *et al.*, 2008), a database of – and ontology for – Chemical Entities of Biological Interest. The ChEBI ontology provides the basis for data collection in this work, as shown in figure 2.1 in chapter 2. In terms of three-dimensional conformations of protein structures, the world wide Protein Data Bank (wwPDB , Berman *et al.*, 2007) is the main repository for experimentally derived structures of proteins and other biomolecules. The ligands contained in those structures are identified by a three-letter code. Information about each of these ligands is documented for example in the PDBeChem database, one of the databases provided by the PDBe (Protein Data Bank in Europe Boutselakis *et al.*, 2003). An overview page for each PDB entry including the key literature reference, a number of structural analysis plots and information about the protein, ligands, protein-protein interactions as well as links to other relevant databases is provided by PDBsum (Laskowski, 2009).

In addition to the aforementioned resources, there are numerous databases for many aspects of three-dimensional structures. PDB-Ligand (Shin & Cho, 2005) offers a ligand-centric view of small molecular ligands in the PDB. According to the publication, the database provides the means to browse, superimpose, classify and visualize the ligands. However, PDB-Ligand does not seem to be available at the published Internet address anymore. Het-PDB Navi (Yamaguchi *et al.*, 2004) is another web interface that provides information about which small molecules bind to which PDB structures.

BindingDB (Liu *et al.*, 2007) and PDBbind (Wang *et al.*, 2005) hold information on protein-ligand binding affinities, which have been collected from the literature. The focus of these projects is on drug targets or candidate drugs with published structural data.

Apart from the active site of an enzyme, proteins may have other functional sites, including post-translationally modified sites, ligand binding sites or protein-protein/DNA/RNA interaction sites. Some of these sites are annotated in the PDB files using the PDB SITE records. The PDBSite (Ivanisenko *et al.*, 2005) database provides information about all these functional sites, as well as the interaction sites that were inferred from contact residues. MSDsite (Golovin *et al.*, 2005) provides all the contact residues of each PDB ligand from the biological assembly. The data may be queried with Prosite patterns and short sequences. The service is now called PDBeMotif.

Bioactive or drug-like small molecules are provided by the ChEMBL (Bender, 2010) database. It contains two-dimensional structure information as well as calculated molecular properties and abstracted bioactivities. DrugBank (Knox *et al.*, 2011) holds manually curated information about drugs and drug targets, including ontology and structural information as well as knowledge about function and metabolism. Finally, MMsINC (Masciocchi *et al.*, 2009) holds non-redundant and highly curated information about biomedically relevant chemical structures as well as computationally predicted molecular properties.

1.4.2.2 Enzyme resources

A plethora of resources has been published for the enzyme research community. Similarly to the compound resources, a full review of all available resources is beyond the scope of this thesis.

There are many aspects of an enzyme that are of interest to researchers. BRENDA (Schomburg *et al.*, 2002) provides biochemical and biophysical data. Catalytic sites are available from the Catalytic Site Atlas (CSA, Porter *et al.*, 2004), where either a manually curated core data set or a larger one that contains homologous enzymes, which have been determined by automated homology searches, can be retrieved. The overall reaction that an enzyme catalyses is classified by the Enzyme Commission (NC-IUBMB & Webb, 1992). The Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) maintains this hierarchical four-level classification and the respective data can be obtained from their web site (NC-IUBMB, 2011), IntEnz (Fleischmann *et al.*, 2004) or KEGG ENZYME (Kanehisa & Goto, 2000). The latter two databases provide the E.C. information in a computer-readable way and often organise the information contained in the original E.C. comment field, e.g. about cofactors, into categories.

Research on how enzyme catalysis works is based on the knowledge of enzyme mechanisms. In contrast to the mere overall reaction (as classified by the E.C.), which states the substrates and products of an enzyme reaction, the mechanism of an enzyme describes each step on the reaction coordinate, including all intermediates. Thus, the mechanism describes how an enzyme transforms substrates into products. This information can, to date, not be generated automatically. Hence, curators manually extract the relevant knowledge from the scientific literature and provide an organised view as entries in databases. Mechanistic information is available from BioPath (Reitz *et al.*, 2004), EzCatDB (Nagano, 2005), the SFLD (Pegg *et al.*, 2006) and MACiE (Holliday, 2009; Holliday *et al.*, 2005, 2007a). Catalytic metal ions are characterised, for instance, in Metal MACiE (Andreini *et al.*, 2009).

1.5 Structural data from X-ray crystallography

Although no experimental structure determination has been undertaken for this work, protein structures from the PDB have been analysed. Thus, it is important to understand the methods used to derive the data in order to be able to assess the significance and limitations of the conclusions drawn.

1.5.1 The protein folding problem

Undoubtedly, proteins are crucial entities for all living cells. In addition to purely structural proteins like those involved in the cytoskeleton, membrane-spanning ion and small molecule channels, or signaling proteins, enzymes are an important variety of proteins because they act as biological catalysts. Most of the chemical reactions that a cell requires for survival are very unlikely to occur in a controlled and regulated manner without a catalyst. Therefore, enzymes constitute a catalytic entity that makes controlled reactions kinetically feasible and that provides a point of regulation of these reactions.

Proteins are chains of amino acids that commonly[1] adopt a distinct three-dimensional structure in which they are active. In enzymes, some of these amino acids are involved in binding ligands (substrates, intermediates, products, organic cofactors, metal cofactors or allosteric regulators) and some are actively involved in catalysis by interacting with the substrate, intermediate or product of the

[1] except for intrinsically unstructured proteins or protein regions

Experimental method	**# Structures**	**# with ligands**	**%age with ligands**
X-ray crystallography	57,687	45,355	78.62%
Nuclear magnetic resonance	7,611	1,340	17.61%
Electron microscopy	314	56	17.83%
Hybrid methods	24	15	62.50%
Any other methods	60	22	36.67%

Table 1.2: Overview of the number of protein containing structures deposited in the Protein Data Bank (released on 22-Sep-2010), the number of structures containing ligands, and the resulting percentage, solved by the different experimental methods.

reaction that this enzyme catalyses. The exact positioning of these amino acids is crucial for the reaction to occur, and hence the three-dimensional folding of an enzyme, determines its function.

In theory it should be possible to compute the three-dimensional conformation that an amino acid chain folds into under physiological conditions by computing the minimum energy of all possible conformations of the protein. In practice, however, the three-dimensional structure of a protein cannot be reliably and accurately computed from its amino acid sequence by computational methods because the search space is too large. Although homology-based (e.g. Soding *et al.*, 2005) and *ab initio* methods (e.g. Han *et al.*, 1997) exist to *predict* a protein's three-dimensional structure from its amino acid sequence, there is no known way to accurately solve this computational problem in an efficient way. The protein folding problem is what computer scientists call an NP (Non-deterministic Polynomial time) problem (e.g. Hart & Istrail, 1997, Berger & Leighton, 1998). It means that no algorithm is known, with which computers can efficiently find the optimal solution to this problem in a reasonable amount of time. Therefore, the only reliable way to determine a protein's three-dimensional structure is by undertaking direct experiments.

In table 1.2, the number of PDB entries for each of the different experimental methods has been documented. Please note that only X-ray crystallography has produced enough entries with sufficient reliability to be considered here. Thus, all the protein and ligand structures from the PDB, which have been used throughout this work, have been determined by X-ray crystallography.

1.5.2 X-ray crystallography: method summary

Individual proteins are too small to be resolved by light microscopy. Visible light has a wavelength (λ) between 400 and $700nm$, while the distance between two atoms is approximately $0.15nm = 1.5\text{Å}$ (Rhodes, 2006, 9). The resolution of optical microscopes is limited to $\approx \lambda/2$, which means that visual light is not suitable for protein structure determination. The wavelength of electron microscopes is well suited for biological studies. However, electrons interact very strongly with biological specimens and cause severe radiation damage, which limits the resolution to about $1nm$. X-rays have a wavelength of about $0.1nm$ and their interaction with biological samples is very weak, which makes them appropriate for determining protein structures with atomic resolution.

The diffraction of X-rays caused by a single protein molecule is too weak to be measured (Rhodes, 2006, 9). Therefore, protein crystals are used for X-ray structure determination to amplify the signal. A protein crystal contains many copies of the molecule neatly arranged in a highly ordered regular three-dimensional array or crystal lattice. The unit cell is the smallest and simplest volume element that is completely representative of the whole crystal (Rhodes, 2006, 10), *i.e.* that may be used to build up the whole crystal lattice by translation along the symmetry axes of the crystal. The result of an X-ray experiment is a pattern of diffracted X-rays or reflections, each with a different intensity, which can be used to infer the electron density content of the time and space averaged unit cell.

The electron density in the unit cell can be calculated using a mathematical function (the Fourier transform) from the structure factors, which can be written as functions as well. However, one of the parameters of the latter function, namely the phases, cannot readily be obtained from this experiment. Instead, further procedures must usually be undertaken to estimate and iteratively improve the phases. Once the electron density has been calculated, a model of the molecule is built, based on prior knowledge about the protein, such as its amino acid sequence, protein secondary structure, and amenable biochemical properties (Rhodes, 2006, 28-30).

The three-dimensional representation of the protein may be displayed in a molecular structure viewer as a model that was created by the crystallographer to be chemically realistic and to match the observed electron density as precisely as possible.

1.5.2.1 Quality and resolution

It is important to acknowledge that the protein model that is viewed in the molecule viewer is an interpretation of experimental data and therefore might contain errors or certain degrees of uncertainty. In PDB files, the position of atoms is specified with a precision of three decimal places. However, considering the precision limits of the method (resolution) and the error introduced by the model, this apparent coordinate precision is misleading.

As mentioned in the previous section, the process of obtaining the electron density function from the experimental data is difficult and comprises steps of estimation and the iterative refinement of the model and the phases.

Several statistical values have been derived and are now commonly used to judge the quality of crystallographic data and the refinement process. The *R-factor* (Stout & Jensen, 1989) is a quality measure that compares the difference of the measured data (structure factors) with the expected experimental data, back calculated from the final atomic model (corresponding to its electron density function), assuming that it was 100% correct. If all of the experimental data is used during the iterative refinement of the model building, the systematic error may be underestimated. Thus, the R_{free}-factor (Brunger, 1992) is a better measure of quality, as it excludes a subset of the data from being used for the model building process. This subset is then used for computing the difference in measured and expected structure factors. R_{free} is always greater than or equal to R. It is a more robust measure for the quality of a crystal structure and it is stated together with the R-factor for the reflections that have been used for refinement (R$_{work}$). R$_{work}$ is commonly smaller than R$_{free}$, but differences of more than several percent might indicate over-refinement of the model or other issues.

A *B-factor* (Debye-Waller factor, Debye, 1913, Waller, 1923) is calculated for each atom in the unit cell. It is a measure for how accurate is the precise location of an atom (RCSB Protein Data Bank, 2011). One component of this inaccuracy is the physical property of all matter that causes an atom to vibrate around a central location. The extent of this movement is temperature-dependant. Thus the B-factor may be referred to as a measure for *thermal vibration*. Sometimes the position of an atom is less accurate than expected from the effect of thermal vibration only, e.g. due to imperfections in the crystal or flexibility in the protein. Even if the reason for this inaccuracy is unknown, it may contribute to the B-factor, which may thus be referred to as the *atomic displacement factor*. Similarly, the *Occupancy* is a number between 0 and 1 that specifies the percentage of the molecules in the crystal, in which a certain atom is located at the given coordinates (RCSB Protein Data Bank, 2011). This allows for alternative locations of highly flexible parts of the protein.

The location of electrons around an atom has a three-dimensional Gaussian probability distribution. The *correlation coefficient* measures the correlation between the shape of the map around the atom and the shape of the expected distribution as calculated from the location of the atom in the model (Brunger *et al.*, 1991; Das *et al.*, 2001). It is a further measure for the quality of a model, which is now widely used in the protein crystal structure community (e.g. Kleywegt *et al.*, 2004).

The resolution of a crystal structure is measured in Ångström (1Å$= 10^{-10}m$) and refers to the minimum distance between two points that can be distinguished. Although there is a large number of quality assessment methods available (the most common ones are mentioned above), resolution is a straightforward and robust parameter to assess the quality of a protein structure model. In this work, a resolution cut-off of 2Å was used to distinguish low and high-resolution structures.

1.5.2.2 Implications for derived data

The side effects of the experimental methods as well as the quality of protein models may impose limitations to the conclusions that can be drawn from the data, which will be discussed in this section.

The position of hydrogen atoms in protein crystal structures are more uncertain and model-dependant than heavy atoms with more electrons (Blakeley *et al.*, 2006; Esposito *et al.*, 2002). Indeed, these positions are often not[1] modelled during the iterative refinement process at all but instead computationally added. Although the quality of these computational methods may be high, it is wise to base scientific research on conformations of molecules on the positions of heavy atoms, rather than the positions of hydrogen atoms.

Furthermore, proteins do not adopt one static conformation but show dynamic motion in solution. The structure determination process yields a time and space averaged snapshot of a protein's three-dimensional structure in the crystalline state, which may lead to the misconception of a protein as a static entity. Although it is important to keep this fact in mind when analysing protein structures, this snapshot does offer valuable information for the viewer, firstly because it allows us to visualise the protein's three-dimensional structure at all and secondly, because the protein adopts a local energy minimum conformation in the crystal, which it very likely often adopts in the cell as well. While the cores of globular proteins are often less dynamic, the solvent-accessible loops on the surface of the proteins may be rather flexible.

Looking at a published structure, one should further keep in mind that the quaternary structure arrangement of identical (homo-oligomers) or different (hetero oligomers) protein chains in the crystal unit cell does not necessarily represent the biologically active arrangement *in vivo* (Krissinel & Henrick, 2007). To distinguish between these quaternary structures, the complex in the crystal is referred to as that in the *(crystallographic) unit cell*, and the biologically active form is called the *biological assembly*.

[1] except for ultra-high resolution structures

1.5.3 Other protein structure determination methods

For the analyses in this work, the focus lies exclusively on X-ray crystallography structures of proteins. The CoFactor database also contains NMR structures in the superposition and solvent accessibility pipelines (see chapter 3).

Table 1.2 shows the different experimental structure determination methods, the number of deposited entries in the PDB and the percentage of those structures that include ligands for each method. In order to work on a homogeneous data set that is as large as possible for statistical purposes, the decision has been taken to only consider protein structures determined by X-ray crystallography for analysis throughout this work.

1.6 Research objectives

In order to widen the understanding of enzyme catalysis, it is necessary to learn about the three catalytic entities in enzymes: catalytic amino acids, metal ions and organic cofactors. In our research group, the former two are already under investigation in context of the MACiE (Holliday *et al.*, 2007a) and Metal MACiE (Andreini *et al.*, 2009) projects. Thus, the aim of this thesis is to investigate organic enzyme cofactors. Objectives include to define which molecules are organic enzyme cofactors and what are their molecular functions and enzymatic mechanisms. The aim is further to determine how organic enzyme cofactors vary in their physicochemical properties, if and how they differ from other metabolites in the cell and if there are intrinsic groupings among them. From a functional perspective, which enzyme reactions depend on organic cofactors will be investigated, how their functional profile differs from that of catalytic amino acids and metal ions and if they can be classified functionally. Finally, the three-dimensional conformation of cofactors will be analysed with respect to its variability when bound to the protein and in solution, as well as the extent of variation in conformations among homologous proteins.

Chapter 2

Methods

In this chapter, the methods employed to produce the results presented in the later chapters are described. This includes the computational methods applied to compile the main data set in the first section, the methods for the data analysis on one- and two-dimensional cofactor data in the second section, and the methods for the three-dimensional structure analysis in the third section.

2.1 Compilation of the data in the CoFactor database

This section describes the data collection process used to populate the CoFactor database (see Chapter 3). It is summarised in figure 2.1. For the definition of what is considered a cofactor in this work, please refer to chapter 1.

The CoFactor data set has been garnered from the publicly available web services of ChEBI (Degtyarenko *et al.*, 2008), PDBsum (Laskowski, 2009), KEGG COMPOUND (Kanehisa & Goto, 2000) and UniProt/SwissProt (Boutet *et al.*, 2007; UniProt Consortium, 2010) as well as the data sets of IntEnz (Fleischmann *et al.*, 2004), MACiE (Holliday *et al.*, 2007a) and PROCOGNATE (Bashton *et al.*, 2006), which have been parsed or queried, directly or using a web service. These resources are used to collect E.C. numbers, CATH codes (Greene *et al.*, 2007), HET identifiers from PDB files, and PDB files associated with each cofactor.

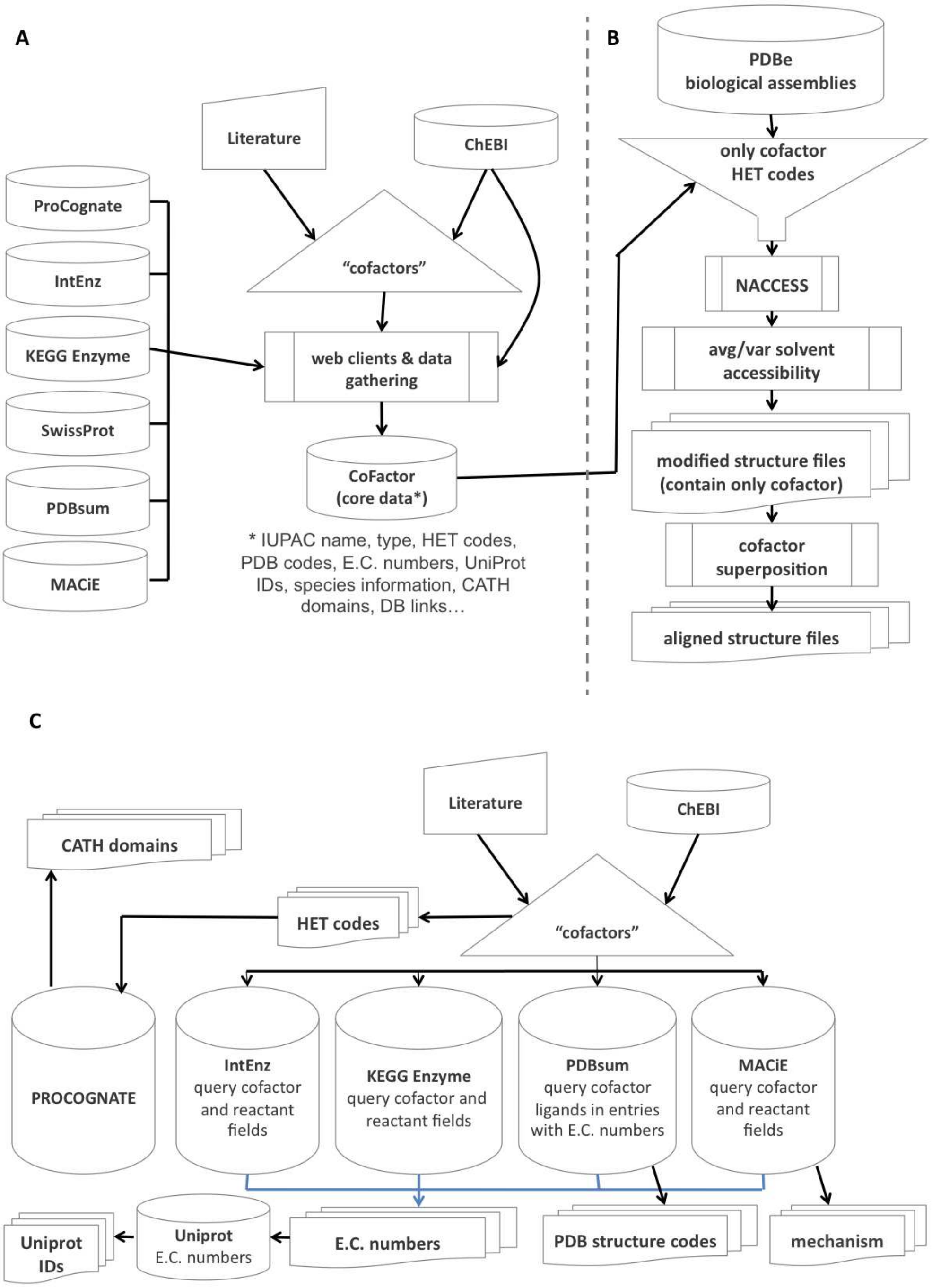

A
Literature
ChEBI
ProCognate
IntEnz
KEGG Enzyme
SwissProt
PDBsum
MACiE
"cofactors"
web clients & data gathering
CoFactor (core data*)
* IUPAC name, type, HET codes, PDB codes, E.C. numbers, UniProt IDs, species information, CATH domains, DB links…
B
PDBe biological assemblies
only cofactor HET codes
NACCESS
avg/var solvent accessibility
modified structure files (contain only cofactor)
cofactor superposition
aligned structure files
C
CATH domains
Literature
ChEBI
HET codes
"cofactors"
PROCOGNATE
IntEnz query cofactor and reactant fields
KEGG Enzyme query cofactor and reactant fields
PDBsum query cofactor ligands in entries with E.C. numbers
MACiE query cofactor and reactant fields
Uniprot IDs
Uniprot E.C. numbers
E.C. numbers
PDB structure codes
mechanism

Figure 2.1: Flow chart of data collection for the CoFactor database. A: Overview, B: Pipeline for solvent accessibility and 3D superposition calculations, C: Detailed view of automated data collection.

Figure 2.1A summarises which databases have been used to generate the CoFactor core data. A more detailed view, which also explains how the data from the different resources map to each other, is shown in figure 2.1C. From ChEBI, the HET[1] codes, systematic molecule names and their relationships to the cofactor, coenzyme or prosthetic group ontology entities (see below) are extracted. As not all cofactors that suffice the working definition were integrated into the ChEBI ontology when this work was started, there is a possibility to manually include these cofactors (see "Literature" unit in figure 2.1C). This list of cofactors is then used to query the cofactor fields of IntEnz, KEGG ENZYME and MACiE, as well as the PDBsum, to obtain a list of E.C. numbers that use each cofactor. As some of the cofactors may transfer functional groups or generally change in the course of the reaction, the substrate and product fields are queried, too, for these cofactors. PDBsum is further used to obtain the PDB codes associated with these enzymes while the mechanistic information is extracted from MACiE. The E.C. numbers are used to query SwissProt in order to obtain a list of biological species that have protein sequences in their genomes, which are known to encode the enzymes associated with the query E.C. number. The list of cofactor HET codes that is obtained from ChEBI is used to query PROCOGNATE for CATH domains that bind each cofactor.

The process of gathering and manipulating the PDBe structure files to obtain the superimposed structures is depicted in figure 2.1B. The coordinate files for all the protein structures that contain at least one of the cofactor HET codes are downloaded from the PDBe biological assembly repository. As explained in section 1.5.2.2, this is a database of quarternary structure arrangements that are more likely to be biologically active than the quarternary structure in the crystallogaphic unit cell. The NACCESS program is employed to calculate the absolute solvent accessibility of all the atoms of the cofactor (identified by its HET code) in the biological assembly. Then, the coordinate files are processed, to obtain modified structure files, which contain only the cofactor (*i.e.* all other coordinates are deleted). To calculate the relative solvent accessibility of each atom in the protein, NACCESS calculates the absolute solvent accessibility of

[1] HET identifiers are 3-letter codes for hetero-atom (i.e. non-amino acid atoms) in PDB structure files.

all the atoms of the cofactor alone (from the modified files, see equation 2.1 and section 2.1.4). The modified structure files are further used to perform the superposition, which is displayed on the CoFactor web pages (see section 2.1.3).

Apart from PDB HET codes and systematic names for its compounds, ChEBI (Degtyarenko *et al.*, 2008) further supplies an ontology of chemical entities of biological interest. That is, ChEBI provides manually curated information on the relationships between the entities in the database. Therefore, ChEBI is used as an entrance point to gather information about which molecules are cofactors in the ontology. Of course, some information in the primary literature was not (yet) represented in ChEBI. To compensate for that, these cofactors were manually included in the search.

Together with information from the primary literature, a data set of 27 cofactors (see table 3.1) was obtained. Many of those are not represented by a single molecule, but by a group of molecules with very similar or identical catalytically active portions (e.g. NAD and NADP).

2.1.1 E.C. numbers of enzyme reactions using cofactors

The UniProtAPI (Patient *et al.*, 2008) is employed to extract species information for each cofactor, based on the E.C. number. This information is considered useful although the author is aware that there may be cases where the enzyme (identified by E.C. number) in one species uses a cofactor whereas in another species it does not. This issue arises from the fact that the E.C. number of an enzyme only classifies the overall reaction catalysed, but does neither imply that all enzymes catalysing this reactions are homologous (*i.e.* arising from a common ancestor), nor that the overall reaction is achieved by the same mechanism. Currently it is not exactly known what percentage of enzymes (defined by E.C. number) have non-homologous instances. Work by Gherardini *et al.* (2007) suggests that such a convergent evolution is non-trivial (ca. 15% on E.C. level 3). In the automatically generated CoFactor data set, however, the comparison takes place on E.C. level 4.

The data collection process uses E.C. numbers, not protein sequences, as a central processing piece (see figures 2.1A and C). This information is taken from the cofactor fields of the IntEnz and KEGG ENZYME databases that provide the information of which E.C. numbers use cofactors. Therefore, in CoFactor, these E.C. numbers will be listed as cofactor-using reactions, even if there are alternative enzymes with the same E.C. number, that may not require this or any cofactor. All of the cofactor-using E.C. numbers are further displayed in the E.C. wheels and E.C. trees (see next section). Therefore, these representations do visualise the functional profile of a cofactor and its distribution over enzyme reaction space, respectively, but this does not imply that these reactions always depend on a cofactor, as they might not, in case alternative enzymes with the same E.C. number exist. The species information, however, may be influenced by this issue, because the E.C. number is used to query SwissProt for protein sequences and the species information, with which these sequences are annotated (see figure 2.1C). Thus, the species list that is displayed on the CoFactor web pages is only a list of all the species that are annotated in SwissProt to encode enzymes with this E.C. number. It does not necessarily imply that all these species have the cofactor-dependant version of the enzyme.

2.1.2 E.C. wheels and E.C. trees

The data collection process (see figure 2.1) yields a list of all E.C. numbers for each cofactor. Each of those E.C. numbers represents an enzyme reaction this cofactor is involved in. To visualise this information, E.C. wheels (Bartlett *et al.*, 2002b) are used (see figure 2.2B). This representation maps the four hierarchical levels of the E.C. classification to four levels in a circular graph, *i.e.* the innermost circle represents the first digit of the E.C. number, the next two rings, from the inside to the outside of the figure, represent the second and the third level of the E.C. classification, *i.e.* sub-class and sub-sub-class. At each level, the angle of each segment corresponds to the percentage of E.C. numbers, and the E.C. number is stated in the label. These wheels allow us to visualise the different types of chemistry that a cofactor is involved in with respect to the E.C. classes it belongs to. An E.C. wheel can be seen as a profile of the overall reaction chemistry of a group of enzymes, in this case all enzymes employing a certain organic cofactor. The script to generate the E.C. wheels was kindly provided by A. Michie, A. Martin and A. Todd.

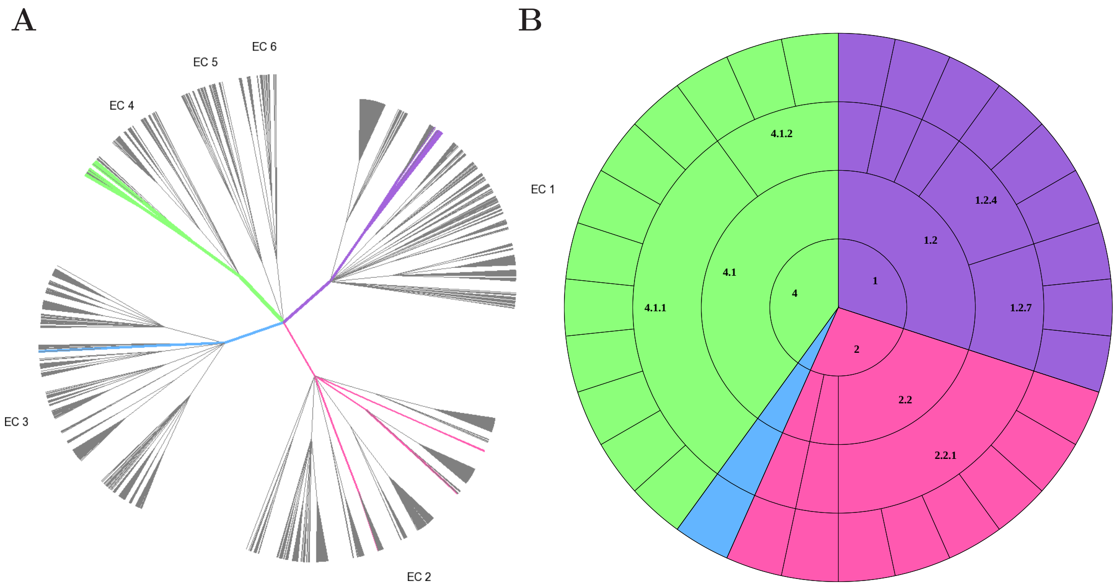


Figure 2.2: Example of an E.C. tree (A) and an E.C. wheel (B).

To visualise how much of enzyme reaction space is occupied by a cofactor, one needs a way to illustrate all of the E.C. numbers currently in use and to highlight those E.C. numbers, by which the cofactor in question is employed. The ec_tree.py program (Dr. N. Furnham, 2009, unpublished) is used to generate an E.C. tree (see figure 2.2A) for each cofactor: an edge in the tree represents one digit of an E.C. number and each leaf is a full four digit E.C. serial number. The tree contains all currently active E.C. numbers for each serial number that employs the cofactor in question. The full branch from the root to the leaf is highlighted in the colour specific for each of the six E.C. classes.

2.1.3 Three-dimensional conformational variability of each cofactor in the PDB

For the cofactor superposition and solvent accessibility computations, all the cofactors (based on HET codes) from the PDBe biological assembly database (Boutselakis *et al.*, 2003) were chosen. This source of structures was used because the quaternary structure (*i.e.* the spatial chain arrangement) of a protein often differs from the coordinates provided in the standard PDB files. While the latter is shown as the unit cell (see section 1.5.2.2) the PDBe provides chain assemblies, which are more likely to be biologically realistic. In the case that an active site is located at the interface of two or more chains and that at least one of these chains in the unit cell is not in the same position as in the biological assembly, it is possible that a cofactor adopts a different conformation in vivo than in the PDBe biological assemblies. However, the complex in the crystal will be a low energy conformation and will usually respect the active arrangement. In any case, the PDBe biological assemblies are considered the best suitable data available. The whole process of preparing these data for the analysis is summarised in figure 2.1B.

In order to illustrate the conformational space that a cofactor occupies when bound to a protein, all suitable instances of this cofactor (as defined by PDB HET identifier) are superimposed, using the same structure files mentioned in section 2.1.3. The most rigid part of each cofactor (aromatic ring structures where possible) has been selected manually and all structures of the same cofactor are superimposed on these atoms. Not all atoms are used for this superposition, but only the most rigid part, in order to optimise the figure to improve understanding by visual inspection. The resulting structure ensembles have been clustered with the KGS method (Kelley *et al.*, 1996), which has been developed to cluster NMR ensembles based on the root mean square deviation (RMSD) into spatially similar clusters. This method finds the number of clusters that minimises the spread within the clusters while maximising the population density of the clusters. In the CoFacotor database, the structures on the superposition page are coloured by cluster membership.

It is worth noting that the superposition on a rigid portion of the molecules is necessary for an easy-to-understand visual representation, but does not provide the best measure to compare the conformations of the cofactors. The minimum RMSD (see section 2.3.1) between all atoms of two structures is used to measure the conformational variability numerically.

2.1.4 Measuring the variation in solvent accessibility of each cofactor in the PDB

All X-ray and NMR structures in the PDBe biological assemblies database (Boutselakis *et al.*, 2003), which contain a HET group assigned to a cofactor, have been used for the superposition and solvent accessibility calculations. NACCESS (Hubbard & Thornton, 1993) V2.2.1 was applied to each structure to compute the solvent accessibility of each atom a in each cofactor twice: first for the biological assembly ($SA_{biolAssembly}(a)$) and second for the cofactor alone $SA_{cofactorAlone}(a)$. The relative solvent accessibility $RSA(a)$ of each atom a has been calculated as shown in equation 2.1.

$$0 < RSA(a) = \frac{SA_{biolAssembly}(a)}{SA_{cofactorAlone}(a)} < 1 \tag{2.1}$$

The relative solvent accessibility of a molecule in a protein structure is the percentage of the surface area of this molecule that is accessible to a solvent. It is calculated by summing over the $RSA(a)$ for all atoms in the molecule. This property may be used to assess the variability of the way a cofactor is bound to proteins. These data are further used to compute the average RSA and 2.1B shows the work flow that produces these data.

2.1.5 Stepwise mechanisms

The mechanism diagrams shown on the CoFactor mechanism pages are based on all the information on a cofactor molecule in MACiE (Holliday *et al.*, 2007a), which have been visually inspected. All substrates and products have been abstracted to be reduced to the essential bonds that are involved in the reaction mechanism catalysed by the respective cofactor.

2.2 Data analysis methods for descriptor variables and two-dimensional chemical structures

In this section, the methods used to produce the results in chapter 4 are discussed.

2.2.1 Compilation of the cofactor and metabolite data sets for the physicochemical properties analysis

2.2.1.1 The cofactor data set

A non-redundant set of PDB HET codes that represent the organic enzyme cofactors has been extracted from the CoFactor database (see chapter 3, Fischer *et al.*, 2010b). Some cofactors are represented by several molecules with different PDB HET identifiers. To reduce redundancy, only one of these HET codes is used to represent the cofactor in case there are several HET codes for molecules representing different states (e.g. protonated and unprotonated in NAD and NAI, or with and without the group they transfer in PLP and PMP) of the same cofactor, but all of the HET codes are used, if they differ in heavy atoms, other than the group they transfer (e.g. MTE and PTE). The resulting data set comprises thirty molecules, which are shown in table 4.1. This excludes different states of the same cofactor but includes different cofactors even if they are similar to each other.

2.2.1.2 The metabolite data set

To construct a comprehensive metabolite data set, all the molecules in KEGG COMPOUND (Kanehisa & Goto, 2000) have been used. Those molecules that do not have a two-dimensional structure file associated with them (e.g. generic entries such as “Hydrogen-donor”) and those that have a placeholder atom “R” or an “n” times repeat of a substructure, have been filtered out because these molecules are generic, and hence the property values (see section 2.2.2) cannot be calculated for them. After filtering, this results in 12,548 molecules. Although undoubtedly incomplete (since many reactions remain unknown), this "best-available" data set is used to represent the chemical space of metabolites in living cells.

2.2.2 Computing physicochemical descriptors of organic compounds

In order to get an overview of the physicochemical properties of organic cofactors and other metabolites, the physicochemical descriptor variables described in this section have been calculated for the cofactor and the metabolite data sets. First, however, all the molecules in the data sets were normalised, as described below.

2.2.2.1 Normalisation of organic molecules

Organic molecule-representing structure files may be stored as a two-dimensional or three-dimensional structure. Further, some programs explicitly store the position of the hydrogen atoms in the file whereas others only store the position of the heavy atoms and rely on the interpreting program to calculate the position of the hydrogen atoms. In order to ensure that all the molecules in the two data sets are represented in two dimensions and stored without explicit hydrogens, a program using CDK (Steinbeck *et al.*, 2003, 2006) was written.

2.2.2.2 Choice of molecular descriptor variables

The following molecular descriptors were selected to be calculated for the cofactor (and indeed the metabolite) data set:

1. **Molecular weight (mw):** the molecular weight of an organic compound is the first obvious choice as it measures the molecular mass, and thus the size of the compound. The OpenBabel program obprop (Fontaine, 2009; Hutchison *et al.*, 2009) was used to calculate the exact molecular mass for each molecule.

2. **Polar surface area (PSA):** the polar surface area measures the polarity of the compound, which may be important for its binding properties to the enzyme active site as well as for its catalytic properties. The polar surface area is the area of the surface of a molecule which is polar and it is measured in $\mathring{A}^2$. PSA is also calculated with obprop.

3. **LogP value (logP):** the logP value is a measure of the hydrophobicity in an organic compound. It can be experimentally determined by measuring the partition of that compound between a hydrophobic (usually n-octanol, $c_{octanol}$) and a hydrophilic (usually water, c_{water}) phase and then calculating the logarithm of their ratio (see equation 2.2). Consequently, a high logP value indicates a more hydrophobic compound. Computationally, logP values can be predicted by summing over the experimentally determined and stored substituents' logP values since the contributions are additive (Ghose *et al.*, 1998). Again, obprob is used to calculate the logP values.

4. **Number of rings (numRing):** The number of rings in an organic molecule is a measure of its rigidity as well as a crude measure of hydrophobicity, given that rings in biological molecules are often aromatic. The obprop program is used to calculate this value for the data sets.

5. **Number of hydrogen bond acceptors (HBA):** the number of hydrogen bond acceptors (calculated by CDK) is crucial for a compound's interaction with the protein and the substrates, and may thus be an important parameter for its function.

6. **Number of hydrogen bond donors (HBD):** similarly, the number of hydrogen bond donors (calculated by CDK) also determines the compound's binding and catalytic behaviour inside the host enzyme.

7. **Percentage of rotatable bonds (percRotBd):** The number of rotatable bonds in an organic molecule determines its flexibility. To normalise this number, it is divided by the total number of bonds to obtain a percentage. This eliminates the size factor from this variable.

8. **Atomic composition (atomComp):** Finally, the atomic composition of an organic molecule is defined as the percentage of polar heavy atoms in all heavy atoms. This is a measure for polarity and calculated using CDK following equation 2.3.

$$logP = \frac{c_{octanol}}{c_{water}} \tag{2.2}$$

$$atomComp = \frac{\#N + \#O + \#S + \#P}{\#N + \#O + \#S + \#P + \#C} \tag{2.3}$$

2.2.3 The *t*-test

A *t*-test is a statistical method that may be used to compare the means of two populations (distributions) (Bremer & Doerge, 2010, 78). The hypothesis to be tested is that the two means are equal. The population sizes do not need to be equal (Holmes *et al.*, 2006, 187). The *t*-test has been reported to be robust even for sample sizes as small as 5 (Bremer & Doerge, 2010, 79). The two populations are assumed to be normally distributed and to have equal variance (Bremer & Doerge, 2010, 80),(Holmes *et al.*, 2006, 187).

For this work, the method t.test() of the R-package “stats” (R Development Core Team, 2008) has been used. The test was run as a two-sided *t*-test. The results are considered to be significant for *p*-values ≤ 0.01.

2.2.4 Structure-based chemical similarity

The SMSD program (Rahman *et al.*, 2009) is used to calculate the pairwise similarity of all the two-dimensional cofactor molecules in the data set. This program employs a maximum-common-subgraph method to identify the most similar parts of the two molecules and to consequently assign a similarity score. Each molecule’s two-dimensional chemical structure is represented as a graph, where the nodes of the graph are the atoms and the bonds are the edges. SMSD provides a switch for the consideration of the bond-order: if the similarity distinguishes between single and double bonds it is referred to as a *bond-sensitive* procedure, and if it does not distinguish the type of bond, it is called *bond-insensitive*.

2.2.5 Principal component analysis

Principal component analysis (PCA) transforms v variables measured on o objects into $k \leq v$ uncorrelated variables (Quinn & Keough, 2002, 443), which span the orthogonal subspace with the highest sample variance (Varoquaux, 2011). Therefore, PCA is a method to determine those components in a multivariate[1] set of data that contribute most to the variability in the data set of several inter-correlated quantitative dependant variables (Abdi & Williams, 2010). This method is often used to reduce the dimensionality of a data table (Quinn & Keough, 2002, 443) by selecting the first (*i.e.* most contributing) n principal components and thus eliminating the redundancy caused by correlated variables with minimum loss of information. Dimension reduction is further useful to allow for human-readable two- or three-dimensional graphical representation of the data (Fielding, 2007, 14). PCA is a means of exploratory data analysis. That is, the method is applied in order to explore and understand the data, rather than to test for a given hypothesis (Fielding, 2007, 12-13). Further information on PCA can be found in many textbooks on multivariate data analysis and PCA (Fielding, 2007; Jolliffe, 2010; Manly, 2004). This section is based on the course materials by Guillet & Rodrigue (2009) unless stated otherwise.

Given a number of objects o, for which a number of variables v was obtained each, these data can be written as an $o \times v$ table. In this work, the objects are organic cofactor molecules identified by their PDB HET code, and the variables are the physicochemical properties, which are calculated for each of the molecules as described in section 2.2.2. The resulting data matrix is shown in table 4.1.

2.2.5.1 Variable requirements and limitations of PCA

In order to fulfil the assumptions made to derive the PCA method, the variables need to be continuous (rather than categorical), and should in theory be normally distributed. On strongly non-Gaussian data, a PCA still de-correlates the data, but there is no guarantee that the directions of maximum variance (the principal components) are the quantities of interest (Varoquaux, 2011).

[1] Multivariate as opposed to uni- and bivariate: for each object multiple (>2) variables are analysed.

> “Many statistical texts explain that multivariate normality is an important assumption for a PCA. However, since the analysis is not used to test statistical hypotheses any violations of the assumptions will only impact on the usefulness of the analysis as a guide to the identification of structure. [. . .] Ultimately the success of the analysis, which is mainly exploratory, should be judged by the usefulness of its interpretation.” (Fielding, 2007, 17)

The Gaussian assumption is often impossible to fulfil for real data. Therefore it is common practise to analyse the variables’ distribution roughly, e.g. by producing boxplots (Guillet & Rodrigue (2009), see also section 2.2.5.2) in order to find outliers, as these have a large impact on the minimisation of the least mean square error procedure, which is part of a PCA. If any outliers are detected, the analysing scientist may manually check if the outlier is a mistake in the data. If a mistake in measurement is likely, the whole object (line) or the whole variable (column) is disregarded and excluded from analysis. If not, the outlier is accepted and analysis proceeds.

If there are missing values (e.g. not available, not a number or infinity), the removal of the respective row or column is obligatory, as the mathematical procedure that a PCA is based on cannot operate on incomplete variables.

Often, the scale in which the variables are measured is different for each variable. In order to guarantee that a variable with a value range of say $[0 : 500]$ is not over-proportionally considered in comparison to a variable with a value range of $[-1 : 1]$, it is crucial to centre (equation 2.4) and scale (equation 2.5) the data, where $\vec{x}$ is the original variable, $\bar{x}$ is the mean value of this variable and $|\vec{x}| = o$ is the number of objects. Please note that the scaling is *not* the default setting in many R-packages and thus has to be set explicitly in the script.

$$\vec{x}_{centred} = \vec{x} - \bar{x} \tag{2.4}$$

$$x_{scaled} = \sqrt{\frac{{x_{centred}}^2}{|\vec{x}_{centred}|}}, \forall x_{centred} \in \vec{x}_{centred} \tag{2.5}$$

PCA only decorrelates linearly correlated variables. Hence, if there are non-linear relationships between variables, this will not be detected. Moreover, PCA assumes that large variances (signal) indicate more interesting dynamics than small variances (noise). However, even if this is doubted, the exploratory benefit of gaining a better understanding of the data by decorrelating it, remains.

As mentioned above, the number of objects is required to be greater than the number of variables and the number of resulting principal components is less or equal to the number of variables. A PCA can be used as a means of reduction in dimensions while maximising the variability in the dimensionally reduced but representative version of the data set. This may be considered successful if, as a rule of thumb, the first three principal components cumulatively account for at least 70% of the total variability in the data (Guillet & Rodrigue, 2009). If the percentage of the total variability accounted for in the first three principal components is very small, this indicates that the number of pairwise independant (or at least merely correlated) variables in the data is greater than three.

2.2.5.2 Boxplots for outlier detection

Boxplots are a non-parametric means to graphically display numerical data. They can be used to get a rough idea about the distribution of the data and possible outliers. Five numbers are calculated from the data to plot the box and whiskers: the minimum value, the lower quartile (Q1), the median (Q2), the upper quartile (Q3) and the maximum value. Figure 2.2.5.2 illustrates how this kind of visualisation works (Guillet & Rodrigue, 2009).

The line in the middle of the box is the median of the data (Q2), whereas the upper and lower bound of the box is the upper and lower quartile (Q3 and Q1), respectively. The median divides the ordered data in two halves, and the upper and lower quartiles divide those two halves again in halves. Thus, the length of the box (interquartile range IQR) represents the middle 50% of the data, and each part of the box represents 25% of the data. The upper whisker is plotted from the upper end of the box (Q3) to the datum with the largest value that is maximum 1.5 times the length of the IQR away from Q3 (and analogously for the lower whisker). Each datum outside the range of the whiskers is plotted as a circle to indicate that it is an outlier.

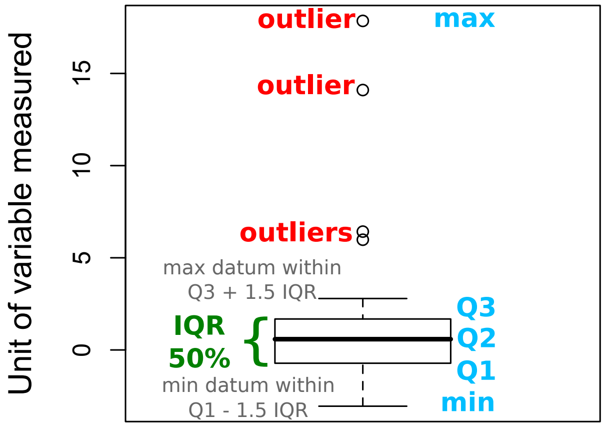


Figure 2.3: An example of a boxplot.

2.2.5.3 PCA: method summary

Once all the preconditions have been checked, the PCA can be performed. To understand the algorithm, one may think of the v variables as axes in a v-dimensional space. If two variables are independant (*i.e.* they do not correlate at all, $cor(v_1, v_2) = 0$), then the angle between these axes is 90°, if they correlate perfectly ($cor(v_1, v_2) = 1$), the angle is 0° (*i.e.* the axes are equivalent), if they anticorrelate the axes are equivalent with opposite directionality ($cor(v_1, v_2) = -1$), and if they partly correlate the angle is somewhere in between.

PCA is a transformation of the data in the original coordinate system defined by its (possibly) interdependant variables, to a Cartesian coordinate system, in which all of the axes are pairwise orthogonal and define the directions of maximum variance in decreasing order. These axes are called the principal components of the data set. The coefficients of the principal components, *i.e.* the contribution of each of the originally measured variables to each of the principal components are called *loadings* (Fielding, 2007, 16). Often, the original variables are shown as vectors (arrows) in a PCA plot. The more a variable contributes to the principal components shown in the plot, the longer is the arrow representing it. Similarly, the more a variable contributes to one of the principal components, the smaller is the angle between this arrow and the axis of the respective principal component. When two arrows have a small angle between them, one can deduce that these variables are highly correlated. Analogously, the objects can be plotted in this new coordinate system of principal components as points, using the same principles. The coordinates of the objects in the new system are called *scores*.

To compute the transformation from dependant variables to principal components, the algorithm calculates the eigenvectors and eigenvalues[1] of the $v \times v$ correlation or covariance matrix of the input data (Fielding, 2007, 19–20). These are then sorted in decreasing value of eigenvalue. The eigenvector with the largest eigenvalue becomes the first principal component. It passes through the mean, minimises the sum squared error with the objects and accounts for most of the variability in the data set. All following principal components contribute less to the variability than the previous principal component.

To sum up, a PCA is a transformation of coordinates from a potentially non-orthogonal system to an orthogonal one such that the first transformed axis represents the greatest variance in the data.

2.2.5.4 PCA: biplot and individuals plot

Useful representations of PCA results include biplots and individuals plots. Biplots show both the objects of the input data matrix as points and the variables as vectors in the space of the chosen first n principal components. This representation enables the viewer not only to see the similarity of the objects to each other and correlations among the variables, but also the contribution of each of the variables to the positioning of the object on the plot. An example can be found in figure 4.5. Of course, all of this only applies for the (usually two) PCs that are plotted.

[1] An eigenvector and an eigenvalue defines the direction and magnitude of a dimension, respectively (Fielding, 2007, 19).

A representation that only shows the objects but not the variables may be required in some cases, for example when the position of different sets of objects is analysed. This plot of "objects" or "individuals" only shows the scores of the PCA, but not the variables. An example can be found in figure 4.8.

2.2.5.5 PCA: correlation circle

Correlation circles visualise the relationships among the variables. The unit sphere of the PCA is a |PCs|-dimensional sphere of radius 1. Each of the descriptor variables can be described as a vector of length 1 that lies on the unit sphere. The projection of these vectors and the unit sphere onto a set of two PCs is the correlation circle of these two PCs, where the circle is the projection of the sphere. The closer a vector is to the circle, the higher is the corresponding variable's contribution to the two chosen PCs. The implementation used in this work is modified from the R-function s.corcircle in package FactoMineR (Husson *et al.*, 2008). Examples of correlation circles can be found in figure 4.7.

2.2.6 Clustering

In Chapter 4, one aim of the analysis is to identify intrinsic groupings in the set of organic cofactor molecules. Generally, there are two types of methods to achieve that: clustering (unsupervised learning) and classification (supervised learning). Both methods aim to group similar items together in groups. The difference is that in classification, there are predefined classes, to which the objects in the data set are assigned, whereas clustering aims to detect both the classes and the class membership of the objects (Fielding, 2007, 2–3,7). Since in the case of organic cofactors there are no known predefined classes, a clustering method is employed to identify the groups of cofactors as well as each molecule's membership. The information in this section is derived from Fielding (2007) and Gan (2007), where further information on the topic may be found.

Clustering methods are algorithms of unsupervised learning and the accuracy of these methods cannot be judged mathematically, but only by the usefulness of the result (Fielding, 2007, 3). They aim to find intrinsic groupings in a data set by analysing pairwise similarity (or dissimilarity) of all data points. Some methods need the original data as input, others may work on a pregenerated distance (or similarity) matrix of the data. Generally speaking, cluster analysis is an unsupervised learning method that assigns a group membership to each data point such that the data points in the same cluster are similar to each other whereas the ones in different clusters are dissimilar.

The Tanimoto similarity measure has been known to yield the best results for chemical similarity clustering of independant pairs of molecules (Willett *et al.*, 1998). Thus, the Tanimoto similarity as calculated by the SMSD program was used for the two-dimensional structure-based clusterings. Equation 2.6 shows how the Tanimoto similarity TS is calculated from the maximum common subgraph representations of the molecules (see section 2.2.4), where c is the size (number of nodes) of the maximum common subgraph between the two compared molecules while a and b are the number of atoms in query and target molecule, respectively. The similarity matrix for all molecules compared against all others is then converted to a distance matrix, such that it can be used by the Ward's clustering method (see next section). For the physicochemical property-based clustering, the result has been calculated directly from the data, such that the R-method *agnes* – agglomerative nesting (hierarchical clustering) from the package *cluster* (Maechler *et al.*, 2005) automatically chooses the default distance metric, which in this case is the Euclidean distance.

$$TS(a,b,c) = \frac{c}{a+b-c} \tag{2.6}$$

There are hierarchical clustering methods and partitional methods. The former may be calculated following a divisive (top-down) or an agglomerative (bottom-up) method. Divisive clustering methods start off with all data points in one cluster (root of the clustering tree, top) and then iteratively divide the most different clusters until each data point is in its own cluster (leaves of the tree, bottom), whereas agglomerative clustering methods operate by starting with all data points in their own clusters and then iteratively merging the most similar clusters until the root at the top contains all data points. Partitional methods may alternatively determine all clusters at once (Fielding, 2007, 47–48).

Various criteria for cluster assignment are published. Common ones are single linkage, average linkage, complete linkage and Ward's criterion. For the purpose of chemical similarity clustering, it has been shown that agglomerative hierarchical clustering using Ward's criterion for cluster assignment yields the best results (Willett *et al.*, 1998). Hence, this method is used for the determination of intrinsic groupings in the cofactor molecules. Ward's criterion iteratively merges those two clusters that produce the smallest increase in the error sum of squares from the mean of each cluster.

2.3 Data analysis methods for three-dimensional chemical structures

This section describes the methods used to obtain the results in chapter 5, where the three-dimensional conformations of organic cofactors are analysed.

2.3.1 Root mean square deviation

When the conformation of ligands in protein structures is analysed, it is crucial to know the experimental quality limits of the data. As elaborated in section 1.5.2, only X-ray crystallographically determined structures are used for analysis and the resolution of the structure must always be considered when interpreting the results. Given that the precision of the coordinates is high enough, the most straightforward measure to compare two spatial conformations of the same ligand is to compute their distance.

$$RMSD_{A,B} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((x_{Ai}-x_{Bi})^2+(y_{Ai}-y_{Bi})^2+(z_{Ai}-z_{Bi})^2)} \tag{2.7}$$

To do this, the root mean square deviation (RMSD) is calculated, which measures the Euclidean distance of the Cartesian coordinates in three-dimensional space (Kabsch, 1976). The RMSD between two molecules A and B is defined in equation 2.7, where both molecules have n atoms with coordinates (x, y, z).

2.3.2 Conformation generation with OMEGA

In this section, the parameters chosen for the conformation generation program OMEGA and the resulting data sets will be discussed, as well as the ligand extraction process of the PDB files and the method used to filter by CATH identifiers.

OMEGA (version 2.4.1, OpenEye Scientific Software, 2004–2010) computationally generates chemically realistic conformations of a given chemical molecule. It first generates models from a fragment library and then performs a torsion angle search on all single bonds. The energy is calculated for each resulting conformer and an energy cut-off (kcal/mol) can be set in the parameters which only allows those conformers in the result set whose difference in energy to the global minimum energy conformer is below the cut-off value.

OMEGA is used here to generate a representative sample of low energy conformations that a molecule would adopt in solution. The program calculates the conformers using the Merck Molecular Force Field ("mmff94s_NoEstat"), which is designed to improve reproduction of aqueous solution phase conformations (OpenEye Scientific Software, 2011).

OMEGA has a wide range of parameters which allow precise adjustment to the individual task it is being used for. Several settings for each of these parameters have been tested for this work and the ones discussed below were found to be most useful.

The SMILES (Daylight Chemical Information Systems, 2008) format, which is a one-dimensional representation of chemical molecules was chosen as the input format. It can store the atom types and their connectivity, as well as stereo chemistry, but bond lengths and angles are not stored explicitly, although the former may be inferred from general chemical knowledge. This one-dimensional representation avoids potential influences from a manually chosen starting conformation. A pipeline program has been written to access the CoFactor database to obtain all the PDB HET codes associated with a given cofactor and the respective ChEBI identifiers. The latter are then used to fetch the SMILES representation of the cofactor from ChEBI. Some cofactors do not have a SMILES representation in ChEBI and OMEGA crashes for some of the SMILES inputs, which results in a reduced data set. The output coordinates were generated in PDB format in order to facilitate the further processing of the output files in the pipeline program.

In order to get an accurate sample of the conformational space in solution, the OMEGA parameters have been chosen to produce evenly spaced conformations which are chemically realistic in the standard force field. The maximum number of generated conformations has been set to 200[1].

Analysis is facilitated if the resulting conformation set is filtered for near-identical conformations. A minimum mutual RMSD parameter can be set in order to guarantee that all conformations in the result set are unique within the chosen limit. Of course the exact value the user might want to chose depends on the properties of the molecule in question. The more rotatable bonds there are in the molecule, the higher the RMSD cut-off. This ensures that OMEGA has a better chance to cover the entire conformational space in solution for highly flexible molecules. OMEGA offers the range increment parameter to partition the input molecules into three groups, depending on the number of rotatable bonds. The value of 7 was chosen here, to obtain 3 groups: 0-6, 7-13 and 14 or more rotatable bonds. The minimum pairwise RMSD values for each of these three groups were set to 0.5Å, 1.0Å and 3.0Å, respectively. The overall energy cut-off was set to 30 kcal/mol for all conformers.

In order to compare the generated conformations to the protein-bound ones from the PDB, the atom naming and numbering was transferred from the PDB files to the generated ones using SMSD (Rahman *et al.*, 2009) although manual intervention was necessary in some cases.

To minimise the bias[2] introduced by using all PDB structures available, two versions of this analysis have been generated: one that includes all structures available and one in which only one conformer (namely the one with the best resolution) per homologous superfamily (CATH identifier) was chosen. Please note that this analysis may reduce the size of the protein-bound data set enormously.

[1] After testing this parameter with higher numbers up to 20000 it was found that 200 is large enough to produce a representative conformational space while being small enough to allow for clear results at visual inspection in molecule viewers.

[2] There is no reason to assume that protein or enzyme space has been crystallised to evenly represent CATH or E.C. space. Possible biases include a bias towards disease-relevant proteins, proteins relevant in human basic metabolism or proteins that are relatively easy to crystallise.

In order to balance the data set size and the bias problem, both versions are analysed in chapter 5.

2.3.3 Superposition method

All superposition procedures were calculated using the SVDSUP program, which was kindly provided by Dr. R. Najmanovich following an algorithm published by Arun *et al.* (1987). The program superimposes two given input molecules in PDB format based on a chosen set of atom numbers by minimising the RMSD (see section 2.3.1) between the atoms in the given set between the two molecules.

Chapter 3

The CoFactor database

This chapter describes the CoFactor database, which serves as the major data source for this work.

Following the definition in chapter 1, the scientific literature was searched manually and the ChEBI (Degtyarenko *et al.*, 2008) ontology and the IntEnz (Fleischmann *et al.*, 2004) database were consulted in order to compile the list of cofactors shown in table 3.1. This table maps each cofactor's name to its main three-dimensional structure (PDB HET) code.

Once the molecules of relevance had been determined and the collection of scientific knowledge about them was initiated, it became apparent that a way of storing the data and retrieving it as required was needed. To solve this problem, the CoFactor database was created.

This chapter is based on my publication (Fischer *et al.*, 2010b). In addition to the published material, however, the database is described here in more detail. The CoFactor database (version 1.0) is publicly available and can be accessed at http://www.ebi.ac.uk/thornton-srv/databases/CoFactor.

Enzymes are proteins that catalyse the repertoire of chemical reactions found in nature, and as such are vitally important molecules. They are generally composed of the twenty common amino acid residues, but many also require small molecules for catalysis to occur. In some cases, these molecules are involved in regulation or in ensuring the correct folding distant from the active site. However, many are termed cofactors, as they are required in the active site and are directly involved in catalysis. These cofactors may be either metal ions, whose

CoFactor ID	Cofactor name	PDB HET code
1	Thiamine diphosphate, ThDP	TDP
2	Flavin adenine dinucleotide, FAD	FAD
3	Flavin mononucleotide, FMN	FMN
4	Nicotinamide-adenine dinucleotide	NAD
5	Phosphopantetheine	PNS
6	Coenzyme A, CoA	COA
7	Pyridoxal 5'-phosphate	PLP
8	Glutathione	GSH
9	Biotin	BTN
10	Tetrahydrofolic acid	THF
11	Adenosylcobalamin, Vitamin B_{12}	B12
12	Ascorbic acid, Vitamin C	ASC
13	Menaquinone, Vitamin K_2	MQ7
14	Ubiquinone, Coenzyme Q	U10
15	Molybdopterin, Molybdenum cofactor	MTE
16	Biopterin	H4B
17	MIO, 4-methylideneimidazole-5-one cofactor	MDO
18	S-adenosylmethionine, SAM, AdoMet	SAM
19	Factor 430, Cofactor F430, Factor F430	F43
20	Coenzyme M, CoM	COM
21	Coenzyme B, CoB	TP7
22	Heme	HEC
24	Dipyrromethane, DPM	DPM
25	Pyrroloquinoline quinone, PQQ	PQQ
26	Topaquinone	TPQ
27	Orthoquinone residues (LTQ, TTQ, CTQ)	-
28	Lipoic acid	LPA

Table 3.1: Data set of all the organic cofactors in the CoFactor database: CoFactor ID, name(s) and the main PDB HET code. For structures, please refer to table 1.1 or the CoFactor web pages. This HET code is used throughout this work.

involvement in catalysis is handled in Metal-MACiE (Andreini *et al.*, 2009), or small organic molecules, which are described here. In both cases, these cofactors extend and enhance the basic catalytic toolkit of enzymes.

To date, there has been little collation of information on organic cofactors and their functions outside of the primary literature. CoFactor has been designed to remedy this, as MACiE (Holliday *et al.*, 2007a) and Metal-MACiE were designed to collate data on enzyme mechanisms and metal ions in catalysis respectively.

3.1 Data content

The CoFactor database contains 27 entries for organic enzyme cofactors (see table 3.1). On the index page, the user can choose which cofactor entry to view. The left-hand navigation contains links to all the pages described below, as well as to the home page, a contact form and a database statistics page. For each cofactor, the web site provides:

1. Overview page – hand-curated information, mostly from primary literature. This includes general information about the molecule, its chemical properties, and about pathways where appropriate.

2. Mechanism (if available in MACiE) – in the standard curly arrow representation of organic chemistry and an optional textual description.

3. Enzymes and domains – enzyme information is integrated with associated 3D structures from PDBe, PDBsum, (Laskowski, 2009), CATH domains (Greene *et al.*, 2007), MACiE enzyme mechanism, proteins that have been assigned this E.C. number according to UniProt (Boutet *et al.*, 2007) as well as a reference that documents the provenance of the information.

   (a) Enzymes that use this cofactor – including visual representations of the cofactor's distribution over enzyme reaction space and its chemical profile, based on the enzyme classification (NC-IUBMB & Webb, 1992).

   (b) Enzymes that synthesise this cofactor.

   (c) Enzymes that recycle this cofactor (if known and applicable).

   (d) Domains that bind this cofactor, taken from PROCOGNATE. (Bashton *et al.*, 2006)

4. Compound – names and identifiers of the same molecule in ChEBI (Degtyarenko *et al.*, 2008), KEGG COMPOUND (Kanehisa & Goto, 2000), PDBeChem (Boutselakis *et al.*, 2003) and PROCOGNATE (Bashton *et al.*, 2006). For each PDB HET code, the web site provides:

   (a) Conformation of the cofactor – shows the superimposed molecules, as described in section 2.1.3 in a three-dimensional molecule viewer.

   (b) Solvent accessibility – displays the average atomic solvent accessibility and its standard deviation as described in section 2.1.4 for each HET code (PDB identifier for non-amino acid molecules) derived from all the PDB codes associated with this cofactor.

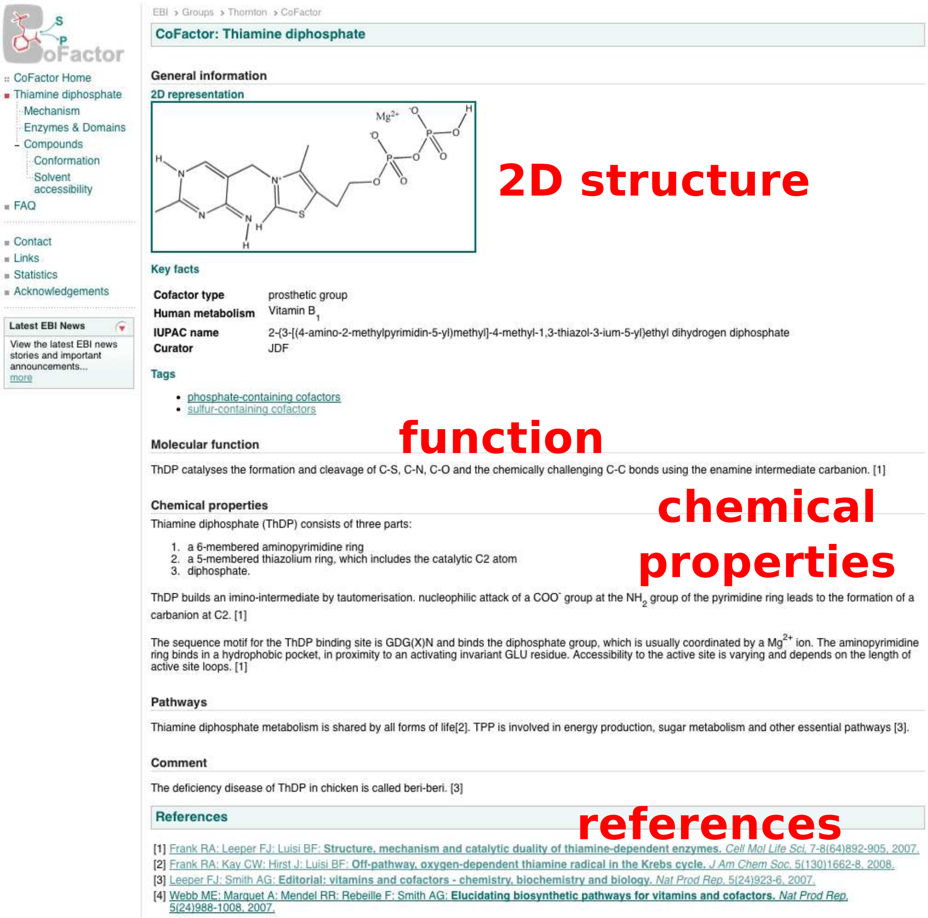
To demonstrate the kind of information stored in the CoFactor database in more detail, here the different pages available for each cofactor are described on the example of thiamine diphosphate.

3.1.1 Overview page

The overview page is separated into at least three sections, followed by the literature references for this page. All the information on this page is manually extracted from the scientific literature. The numbers in square brackets refer to the publication that this specific piece of information has been taken from. A link to the corresponding publication in the CiteXplore database (Literature Services Group, EMBL-EBI, 2011) can be found at the end of the page under the heading "References".

The first section entitled "General information" is compulsory for any cofactor to be included in the CoFactor database. It always consists of a two-dimensional structure of the cofactor molecule and some key facts like the cofactor type according to our hierarchical definition (coenzyme, prosthetic group, both or polypeptide-derived), its relevance in human metabolism (vitamin, biosynthesised or not in humans), the IUPAC name of the molecule, if available from ChEBI, and the initials of the curators involved in this entry. The “General information” section may also display a collection of links if this cofactor has been tagged with a group tag. In the case of thiamine diphosphate, the two groups it is tagged with are phosphate-containing cofactors and sulfur-containing cofactors. Upon clicking, these tags link to a list of all the other entries with this tag.

The second compulsory section is “Molecular function”. It summarises what has been published about the molecular function of this cofactor. This information is backed up by literature references.



EBI › Groups › Thornton › CoFactor

CoFactor Home
- Thiamine diphosphate
  - Mechanism
  - Enzymes & Domains
  - Compounds
    - Conformation
    - Solvent accessibility
- FAQ
- Contact
- Links
- Statistics
- Acknowledgements

Latest EBI News
View the latest EBI news stories and important announcements...
more

CoFactor: Thiamine diphosphate

General information

2D representation

Key facts

Cofactor type	prosthetic group
Human metabolism	Vitamin B_1
IUPAC name	2-{3-[(4-amino-2-methylpyrimidin-5-yl)methyl]-4-methyl-1,3-thiazol-3-ium-5-yl}ethyl dihydrogen diphosphate
Curator	JDF

Tags

- phosphate-containing cofactors
- sulfur-containing cofactors

Molecular function

ThDP catalyses the formation and cleavage of C-S, C-N, C-O and the chemically challenging C-C bonds using the enamine intermediate carbanion. [1]

Chemical properties

Thiamine diphosphate (ThDP) consists of three parts:

1. a 6-membered aminopyrimidine ring
2. a 5-membered thiazolium ring, which includes the catalytic C2 atom
3. diphosphate.

ThDP builds an imino-intermediate by tautomerisation. nucleophilic attack of a COO^- group at the NH_2 group of the pyrimidine ring leads to the formation of a carbanion at C2. [1]

The sequence motif for the ThDP binding site is GDG(X)N and binds the diphosphate group, which is usually coordinated by a Mg^{2+} ion. The aminopyrimidine ring binds in a hydrophobic pocket, in proximity to an activating invariant GLU residue. Accessibility to the active site is varying and depends on the length of active site loops. [1]

Pathways

Thiamine diphosphate metabolism is shared by all forms of life[2]. TPP is involved in energy production, sugar metabolism and other essential pathways [3].

Comment

The deficiency disease of ThDP in chicken is called beri-beri. [3]

References

[1] Frank RA; Leeper FJ; Luisi BF: **Structure, mechanism and catalytic duality of thiamine-dependent enzymes.** *Cell Mol Life Sci*, 7-8(64)892-905, 2007.
[2] Frank RA; Kay CW; Hirst J; Luisi BF: **Off-pathway, oxygen-dependent thiamine radical in the Krebs cycle.** *J Am Chem Soc*, 5(130)1662-8, 2008.
[3] Leeper FJ; Smith AG: **Editorial: vitamins and cofactors - chemistry, biochemistry and biology.** *Nat Prod Rep*, 5(24)923-6, 2007.
[4] Webb ME; Marquet A; Mendel RR; Rebeille F; Smith AG: **Elucidating biosynthetic pathways for vitamins and cofactors.** *Nat Prod Rep*, 5(24)988-1008, 2007.

Figure 3.1: Screen shot of the overview page of the cofactor thiamine diphosphate from the CoFactor database.

The third compulsory section, “Chemical properties”, describes the chemical composition of the cofactor and may contain further information about the molecule's chemical features, such as its binding motif in the protein or its chemical activation reaction, as can be seen for thiamine diphosphate in figure 3.1.

There are two further optional sections. If there is relevant information about the pathways this cofactor is involved in, it can be stated in the "Pathways" section. Relevant information includes the cofactors' involvement in other cofactors' biosynthesis, knowledge about its general or specialised employment in different groups of biological species, or specialised metabolic areas, in which the cofactor is preferentially used.

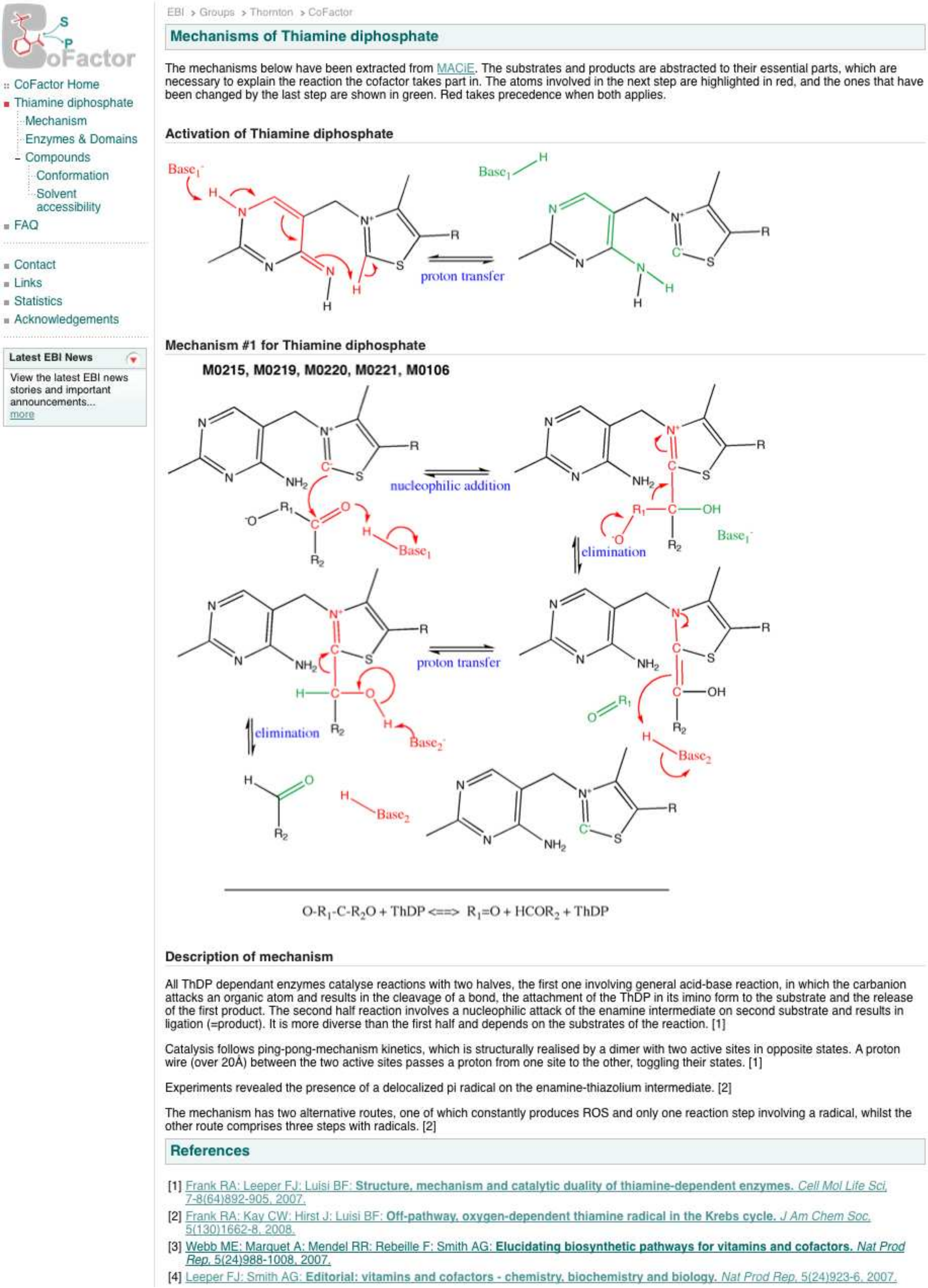
The optional "Comment" section may hold any other information that does not fit any of the above categories (or the mechanism or enzymes it is associated with as these are handled separately). Information about deficiency diseases or genetic disorders associated with this cofactor is often provided here.

As mentioned above, all pages have a literature "References" section, which links to the relevant publications used on this page.

3.1.2 Mechanism page

The mechanism page describes the molecular mechanism(s) this cofactor can perform. The information is mainly based on the data from the MACiE database (Holliday *et al.*, 2007a), which have been reduced and summarised from a cofactor-centric point of view. Where several MACiE mechanisms were available, in which the cofactor performs a chemically identical task, the different substrates and intermediates have been abstracted to the essential bonds involved in the mechanism. The example for cofactor thiamine diphosphate is shown in figure 3.2.

In all depictions of mechanisms, these conventions are followed: in general the two-dimensional chemical structures are shown in black, the atoms that are involved in the following step as well as the curly arrows indicating the reaction chemistry are red, and the atoms that have changed in the previous step are green. If an atom has changed in the previous step and will also change in the following one, then red takes precedence over green. Unless specifically stated in the literature or in MACiE, all reactions are assumed to be reversible. The molecular procedure that happens in each step is annotated in blue under the



View the latest EBI news stories and important announcements... more

Mechanisms of Thiamine diphosphate

The mechanisms below have been extracted from MACiE. The substrates and products are abstracted to their essential parts, which are necessary to explain the reaction the cofactor takes part in. The atoms involved in the next step are highlighted in red, and the ones that have been changed by the last step are shown in green. Red takes precedence when both applies.

Activation of Thiamine diphosphate

Mechanism #1 for Thiamine diphosphate

M0215, M0219, M0220, M0221, M0106

$O\text{-}R_1\text{-}C\text{-}R_2O + ThDP <==> R_1{=}O + HCOR_2 + ThDP$

Description of mechanism

All ThDP dependant enzymes catalyse reactions with two halves, the first one involving general acid-base reaction, in which the carbanion attacks an organic atom and results in the cleavage of a bond, the attachment of the ThDP in its imino form to the substrate and the release of the first product. The second half reaction involves a nucleophilic attack of the enamine intermediate on second substrate and results in ligation (=product). It is more diverse than the first half and depends on the substrates of the reaction. [1]

Catalysis follows ping-pong-mechanism kinetics, which is structurally realised by a dimer with two active sites in opposite states. A proton wire (over 20Å) between the two active sites passes a proton from one site to the other, toggling their states. [1]

Experiments revealed the presence of a delocalized pi radical on the enamine-thiazolium intermediate. [2]

The mechanism has two alternative routes, one of which constantly produces ROS and only one reaction step involving a radical, whilst the other route comprises three steps with radicals. [2]

References

[1] Frank RA: Leeper FJ: Luisi BF: **Structure, mechanism and catalytic duality of thiamine-dependent enzymes.** *Cell Mol Life Sci,* 7-8(64)892-905, 2007.

[2] Frank RA: Kay CW: Hirst J: Luisi BF: **Off-pathway, oxygen-dependent thiamine radical in the Krebs cycle.** *J Am Chem Soc,* 5(130)1662-8, 2008.

[3] Webb ME: Marquet A: Mendel RR: Rebeille F: Smith AG: **Elucidating biosynthetic pathways for vitamins and cofactors.** *Nat Prod Rep,* 5(24)988-1008, 2007.

[4] Leeper FJ: Smith AG: **Editorial: vitamins and cofactors - chemistry, biochemistry and biology.** *Nat Prod Rep,* 5(24)923-6, 2007.

Figure 3.2: Screen shot of the mechanism page of the cofactor thiamine diphosphate from the CoFactor database.

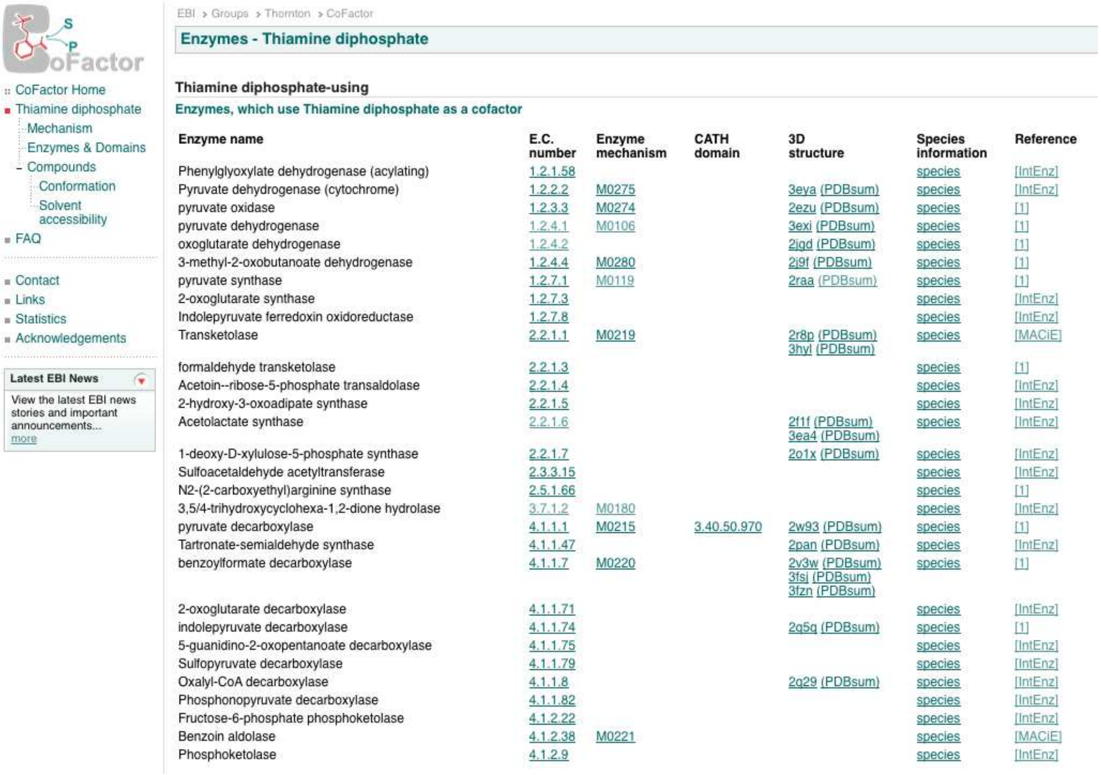
reaction arrow. All mechanisms display the MACiE IDs used to generate them as well as the summary overall transformation at the bottom of the depiction. If there are comments necessary, these will be displayed in magenta.

In case the cofactor has an activation step, the mechanism of activation is shown first. As this is the case for thiamine diphosphate, the carbanionisation is shown first in figure 3.2. Below that, the stepwise mechanism by which thiamine diphosphate breaks carbon-carbon or similarly non-polar bonds is shown. Some cofactors perform two or more distinct mechanisms. If this is the case, all known mechanisms will be shown on the mechanism page. It is worth noting that FAD (cid 2 in CoFactor) is a special case in terms of mechanism: the analysis of the examples in MACiE shows that FAD performs several mechanism "building blocks" that can be combined in various ways. Normally, the cofactor is either restored in its original state (if applicable after activation, e.g. thiamine diphosphate) or transformed to its alternative state (e.g. NAD), in which it leaves the active site to be recycled elsewhere. This is not the case for the FAD mechanistic building blocks. Instead, the cofactor is restored by additional varying reactions that can thus not be generalised or by combining two or more of the building blocks.

The CoFactor curators have an optional field to describe the mechanism(s) above in case there is additional information from the literature, which cannot conveniently be integrated into the depiction, or in case some mechanistic information has been published but it does not suffice to deduce the full mechanism.

3.1.3 Enzymes and domains page

On this page, information about enzymes and domains that use, synthesise, recycle or bind the cofactor is displayed. The section that lists the cofactor-using enzymes has three sub-sections. The first one is a simple list of these enzymes (figure 3.3), identified by their E.C. number (linked to IntEnz), integrated with their mechanism (from MACiE), their three-dimensional structure (from the PDBe and linked to PDBsum), their binding CATH domains (linking to PROCOGNATE), a list of the species known to possess this E.C. number, and finally the source of the reference that states that this enzyme uses the cofactor in question.



Enzymes - Thiamine diphosphate

Thiamine diphosphate-using

Enzymes, which use Thiamine diphosphate as a cofactor

Enzyme name	E.C. number	Enzyme mechanism	CATH domain	3D structure	Species information	Reference
Phenylglyoxylate dehydrogenase (acylating)	1.2.1.58				species	[IntEnz]
Pyruvate dehydrogenase (cytochrome)	1.2.2.2	M0275		3eya (PDBsum)	species	[IntEnz]
pyruvate oxidase	1.2.3.3	M0274		2ezu (PDBsum)	species	[1]
pyruvate dehydrogenase	1.2.4.1	M0106		3exi (PDBsum)	species	[1]
oxoglutarate dehydrogenase	1.2.4.2			2jgd (PDBsum)	species	[1]
3-methyl-2-oxobutanoate dehydrogenase	1.2.4.4	M0280		2j9f (PDBsum)	species	[1]
pyruvate synthase	1.2.7.1	M0119		2raa (PDBsum)	species	[1]
2-oxoglutarate synthase	1.2.7.3				species	[IntEnz]
Indolepyruvate ferredoxin oxidoreductase	1.2.7.8				species	[IntEnz]
Transketolase	2.2.1.1	M0219		2r8p (PDBsum) 3hyl (PDBsum)	species	[MACiE]
formaldehyde transketolase	2.2.1.3				species	[1]
Acetoin--ribose-5-phosphate transaldolase	2.2.1.4				species	[IntEnz]
2-hydroxy-3-oxoadipate synthase	2.2.1.5				species	[IntEnz]
Acetolactate synthase	2.2.1.6			2f1f (PDBsum) 3ea4 (PDBsum)	species	[IntEnz]
1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7			2o1x (PDBsum)	species	[IntEnz]
Sulfoacetaldehyde acetyltransferase	2.3.3.15				species	[IntEnz]
N2-(2-carboxyethyl)arginine synthase	2.5.1.66				species	[1]
3,5/4-trihydroxycyclohexa-1,2-dione hydrolase	3.7.1.2	M0180			species	[IntEnz]
pyruvate decarboxylase	4.1.1.1	M0215	3.40.50.970	2w93 (PDBsum)	species	[1]
Tartronate-semialdehyde synthase	4.1.1.47			2pan (PDBsum)	species	[IntEnz]
benzoylformate decarboxylase	4.1.1.7	M0220		2v3w (PDBsum) 3fsj (PDBsum) 3fzn (PDBsum)	species	[1]
2-oxoglutarate decarboxylase	4.1.1.71				species	[IntEnz]
indolepyruvate decarboxylase	4.1.1.74			2q5q (PDBsum)	species	[1]
5-guanidino-2-oxopentanoate decarboxylase	4.1.1.75				species	[IntEnz]
Sulfopyruvate decarboxylase	4.1.1.79				species	[IntEnz]
Oxalyl-CoA decarboxylase	4.1.1.8			2q29 (PDBsum)	species	[IntEnz]
Phosphonopyruvate decarboxylase	4.1.1.82				species	[IntEnz]
Fructose-6-phosphate phosphoketolase	4.1.2.22				species	[IntEnz]
Benzoin aldolase	4.1.2.38	M0221			species	[MACiE]
Phosphoketolase	4.1.2.9				species	[IntEnz]

Figure 3.3: Screen shot of the cofactor-using list of the enzyme page of the cofactor thiamine diphosphate from the CoFactor database.

To identify, which enzymes employ this cofactor, either information from the scientific literature is used, in which case the reference field will cite a number in square brackets. At the bottom of the page, this number will identify the publication containing this information and link to its CiteXplore (Literature Services Group, EMBL-EBI, 2011) entry. Alternatively, the IntEnz cofactor field or the appearance of known cofactors in the substrate and product fields is used to automatically determine the use of this cofactor in the given E.C. reaction. It is worth noting that in the latter case, it is up to the user to check that the cofactor is actively involved in catalysis in this reaction. A counter example would be the enzyme that synthesises the cofactor. In that case the cofactor will appear on the product side but it does mostly not act as a cofactor in that reaction. In most other cases, however, it is expected that the molecule will indeed be a cofactor when detected automatically using this procedure. Therefore, this small draw-back is accepted in favour of the many reactions that are detected. If a cofactor is annotated as a catalytic cofactor in MACiE, the source will display MACiE instead of IntEnz. This information can later be used to separate manually annotated cases from automatically annotated ones, in order to retrieve high quality data sets.

Further, the mechanistic information from MACiE on the E.C. number is automatically integrated, if available, by cross-referencing the E.C. number in MACiE with the PDB HET atom codes for this cofactor in the MACiE entry. In case such a MACiE entry is detected, it is linked to the corresponding MACiE page.

The CATH domain(s) that bind this cofactor are obtained from the PROCOGNATE database by querying it for the E.C. number. For data analysis, PDBsum was used to detect the binding CATH domains because the work on PROCOGNATE has been discontinued. As this was done after publication of the CoFactor database version 1.0, these data are not yet publicly available but will be published with the next update of CoFactor.

To implement this, a program (see figure 3.4) was written that parses all the binding information from Ligplot (Wallace *et al.*, 1995), which is available for each PDBsum entry. The program filters out all binding interactions that are not between the atoms of the cofactors (identified by the HET code) and the protein. The remaining interactions are then cross-referenced with a data table in the MACiE database that stores the domain borders as defined by CATH for all PDB files. All domains that have at least 3 interactions (independant of the interaction mode defined by Ligplot) are accepted as domains that bind this cofactor. The knowledge about which PDB codes contain the cofactor in question is obtained from PDBsum as well, by automatically parsing the het2pdb.lst file, provided for download by PDBsum. This contains a mapping from PDB HET codes to E.C. numbers. Both, the PDBe (Boutselakis *et al.*, 2003) entry for this PDB code and the PDBsum entry are linked here. Finally, the species link opens a window containing a list of species that have been manually annotated in UniProt (Boutet *et al.*, 2007) to express the enzyme with the given E.C. number. The UniProtAPI (Patient *et al.*, 2008) was used to acquire this information (organism name, NCBI taxon ID and UniProt ID).

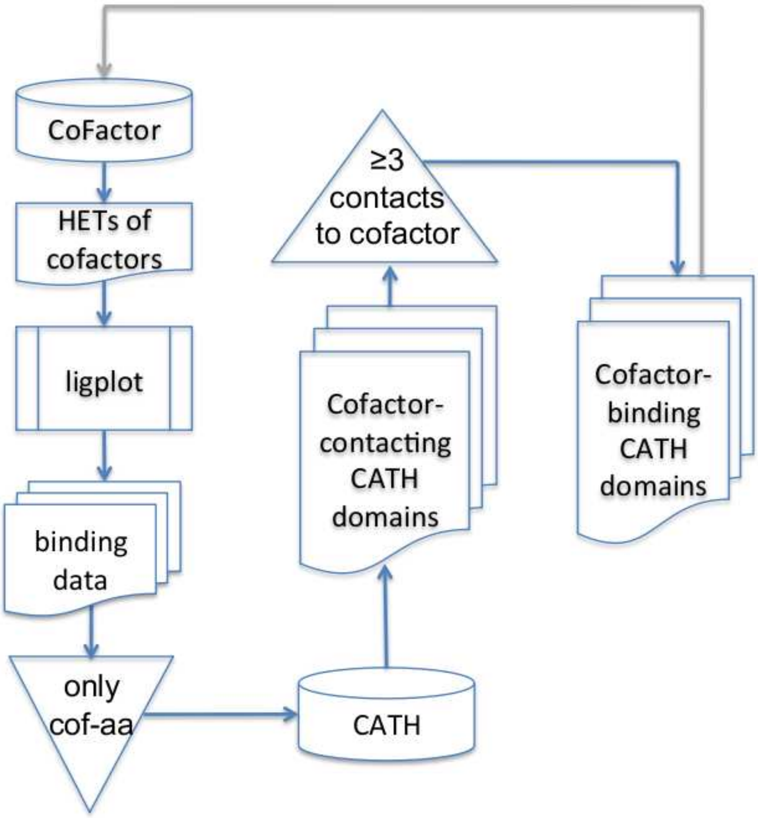


Figure 3.4: Flow chart describing how the CATH domain data is derived from the PDBsum ligplot data.

Bringing all this information about each enzyme together allows the user to get an integrative view of each enzyme that employs the cofactor in question, from protein structure, over domains and species all the way to its detailed stepwise mechanism.

The same information can be visualised in two useful ways: as an E.C. tree or an E.C. wheel, which is shown in the second and third sub-section of the cofactor-using enzymes section. The E.C. tree (figure 3.5) visualises all of enzyme space by showing all current E.C. numbers as a tree (grey), following the E.C. hierarchy. The visualisation script has been kindly provided by Dr. Nick Furnham.

In the context of enzyme cofactor usage, the E.C. tree can be used to display how widespread the distribution of the cofactor is in enzyme space over the six E.C. classes. The E.C. tree highlights the occurrence of this cofactor in each of

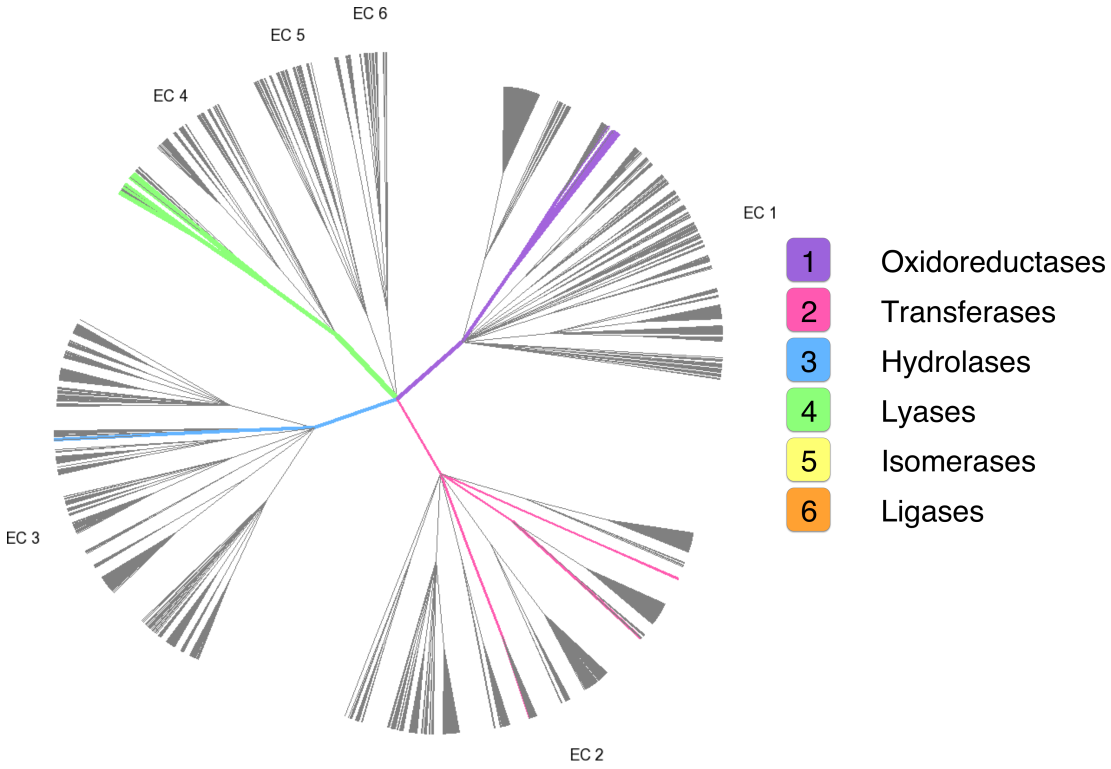


Figure 3.5: Screen shot of the E.C. tree of the enzyme page of the cofactor thiamine diphosphate from the CoFactor database.

the E.C. numbers where the colour of the branch varies depending on which of the six E.C. classes this enzyme belongs to. The oxidoreductases (E.C. class 1) are highlighted in purple, transferases (E.C. class 2) in pink, hydrolases (E.C. class 3) in blue, lyases (E.C. class 4) in green, isomerases (E.C. class 5) in yellow, and ligases (E.C. class 6) in orange. In the example of thiamine diphosphate, one can deduce from the E.C. tree (figure 3.5) that the cofactor is used by a few enzymes in four different E.C. classes, most of which also belong to the same or neighbouring sub- and sub-sub-classes. Only in the transferase class, the localisation of thiamine diphosphate-usage is more widespread over the classification hierarchy. Additionally the E.C. tree shows that there are no known isomerases or ligases that employ thiamine diphosphate.

The E.C. wheels (figure 3.6) use the same underlying information to display a chemical reaction profile of the current cofactor. In comparison to the E.C. tree, it does not quantify the frequency of cofactor employment with respect to all of enzyme space, but it visualises the percentages of E.C. class representation in the overall functional profile of the current cofactor.

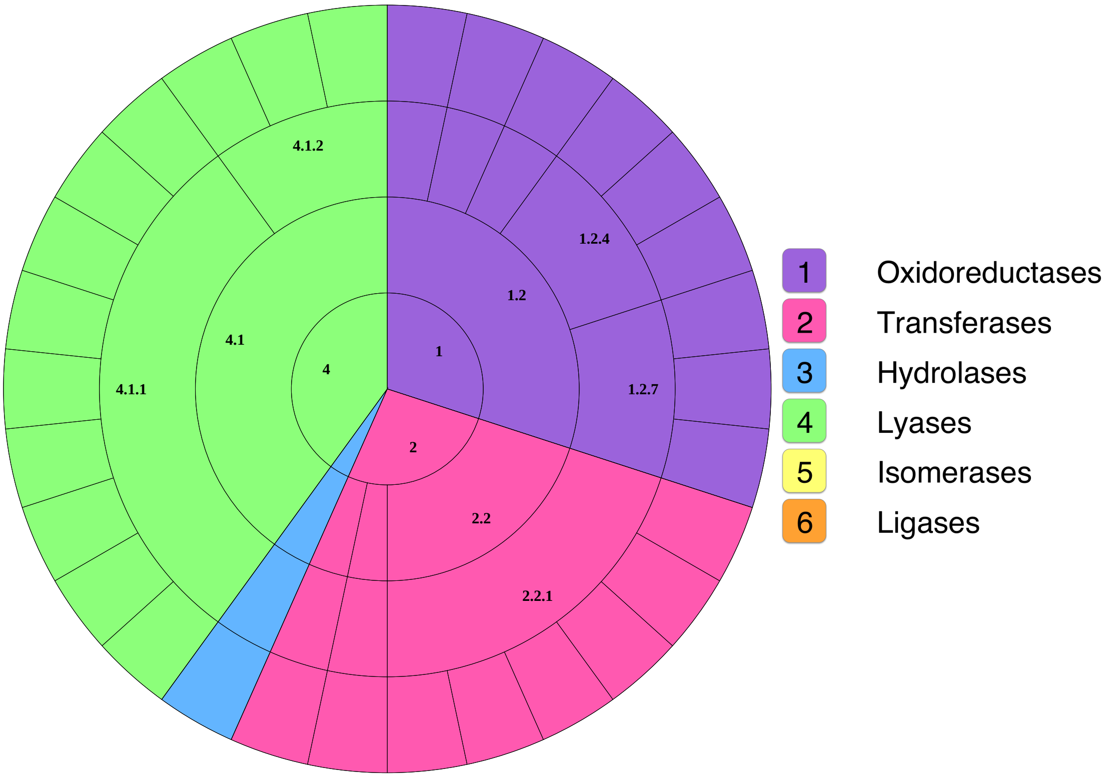
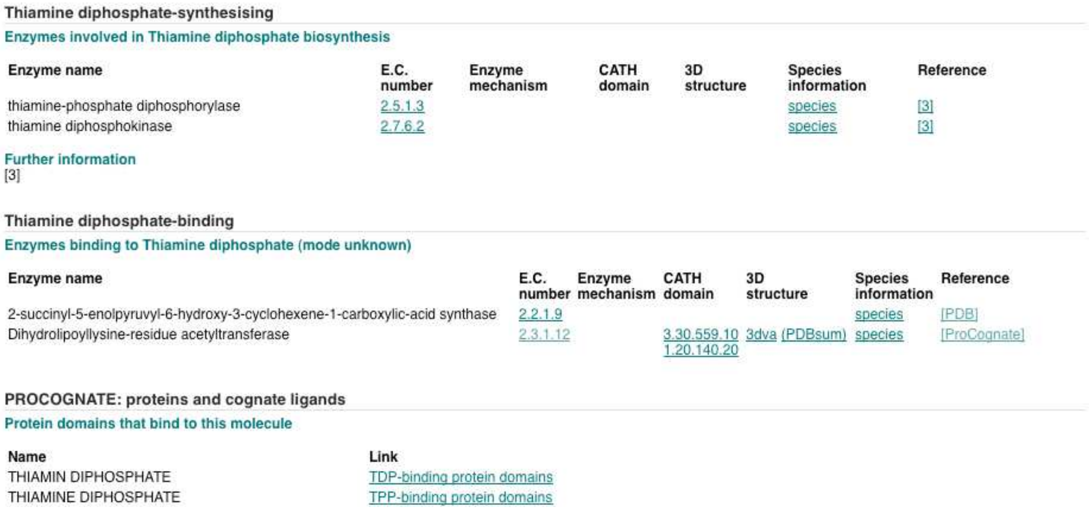


Figure 3.6: Screen shot of the E.C. wheel of the enzyme page of the cofactor thiamine diphosphate from the CoFactor database.

In the example of thiamine diphosphate, the E.C. wheel shows that the reactions that the cofactor enables are approximately one third oxidoreductases, transferases and lyases each. In contrast to the E.C. tree, it does not show the background distribution of those enzyme classes. The colour key is the same is for the E.C. tree.

Together, the list and the two plots provide a thorough picture of the available data on enzymes that use the cofactor of interest. However, there are other relationships between enzymes and cofactors that might be of interest, such as a list of enzymes that are known to participate in a cofactor's biosynthetic pathway (figure 3.7). This list only contains hand-curated information. The table provides the same categories as the list of cofactor-using enzymes above.



Thiamine diphosphate-synthesising

Enzymes involved in Thiamine diphosphate biosynthesis

Enzyme name	E.C. number	Enzyme mechanism	CATH domain	3D structure	Species information	Reference
thiamine-phosphate diphosphorylase	2.5.1.3				species	[3]
thiamine diphosphokinase	2.7.6.2				species	[3]

Further information

[3]

Thiamine diphosphate-binding

Enzymes binding to Thiamine diphosphate (mode unknown)

Enzyme name	E.C. number	Enzyme mechanism	CATH domain	3D structure	Species information	Reference
2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid synthase	2.2.1.9				species	[PDB]
Dihydrolipoyllysine-residue acetyltransferase	2.3.1.12		3.30.559.10 1.20.140.20	3dva (PDBsum)	species	[ProCognate]

PROCOGNATE: proteins and cognate ligands

Protein domains that bind to this molecule

Name	Link
THIAMIN DIPHOSPHATE	TDP-binding protein domains
THIAMINE DIPHOSPHATE	TPP-binding protein domains

Figure 3.7: Screen shot of the cofactor-synthesising list and the domain information of the enzyme page of the cofactor thiamine diphosphate from the CoFactor database.

Sometimes there are data in the PDBe or PROCOGNATE that a certain cofactor HET code binds to an enzyme, but that does not reveal whether it acts as a catalytic cofactor in that enzyme reaction. There are several other reasons one could imagine why the cofactor would bind the enzyme but not catalyse: it could be an allosteric regulator, necessary for correct protein folding but be chemically inert in that protein or it could bind in another (natural) ligand's binding site. To distinguish these cases from the list above, they are displayed as cofactor-binding enzyme with unknown mode.

For cofactors that are not regenerated in each catalytic cycle (mostly coenzymes like e.g. NAD), there can be a further sub-heading listing cofactor-recycling reactions. Like the biosynthesis list, this list does not contain automatically annotated data, but only literature-based manually annotated information.

Lastly, the enzymes and domains page links to the relevant PROCOGNATE pages for each PDB HET code that is associated with the cofactor in question.

In the case of thiamine diphosphate, these are TDP and TPP. The linked page in PROCOGNATE lists the CATH codes and PDB files that bind this HET code.

3.1.4 Compound page: conformation and solvent accessibility

The compound page links out to the four compound-based databases ChEBI, KEGG COMPOUND, PDBeChem and PROCOGNATE. These data have been gathered by hand using the ChEBI database links and the KEGG COMPOUND web search. The two subordinated pages “Conformation” and “Solvent Accessibility” are the analytical pages.

For each of the PDB HET codes associated with this cofactor, the conformation page shows a superposition of all instances of this molecule (maximum one per PDB code) in the PDBe (see sections 2.3.3 and 2.1.3). To ensure the most informative format for the user, the cofactors are superimposed on one functional or structural group, rather than on all atoms. The superimposed portion is the functional portion where possible (and useful for display) or a flat or ring-like portion otherwise. The superimposed structures are displayed in a three-dimensional molecule viewer (Hanson, 2010) applet (Jmol) so that the user can rotate the molecules and even use RASMOL (Sayle & Milner-White, 1995) script commands on the applet’s command line option.

Furthermore the average and variance of the relative solvent accessibility (see section 2.1.4) have been computed and also visualised in a three-dimensional molecule viewer. How these data are generated is shown in figure 2.1.

3.2 The CoFactor data set

In this section, information from the CoFactor overview (see section 3.1.1) and mechanism pages (see section 3.1.2) for each of the 27 cofactors in the CoFactor database is provided. The data have been manually extracted from the scientific literature. For depictions of the two-dimensional chemical structures of the cofactors, please refer to the CoFactor database or to table 1.1.

3.2.1 Thiamine diphosphate

Thiamine diphosphate (ThDP) is a prosthetic group in enzymes. In human metabolism, ThDP is the water soluble vitamin B_1. Its molecular function is to catalyse the formation and cleavage of bonds between heavy atoms, including C-S, C-N, C-O, but also the chemically challenging C-C bonds. An important step in its mechanism is a carbanion intermediate (Frank *et al.*, 2007).

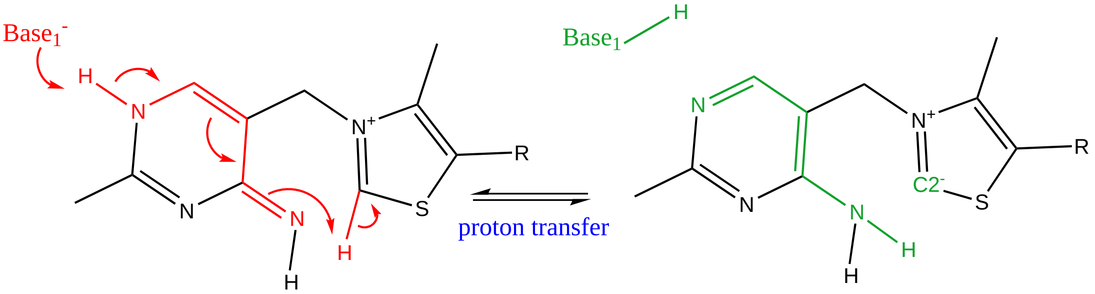


Figure 3.8: Activation mechanism of ThDP. The catalytic C2 atom is marked in green.

The ThDP molecule consists of three portions: a 6-membered aminopyrimidine ring, a 5-membered thiazolium ring (which includes the catalytic C2 atom), and a diphosphate group. The cofactor builds an imino-intermediate by tautomerisation. The nucleophilic attack of a COO^- group at the NH group of the pyrimidine ring leads to the formation of the catalytic carbanion at C2 (see figure 3.8). The sequence motif for the ThDP binding site is GDG(X)N and binds the diphosphate group, which is usually coordinated by a Mg^{2+} ion. The aminopyrimidine ring binds in a hydrophobic pocket, in proximity to an activating invariant glutamate residue[1]. Accessibility to the active site is variable and depends on the length of active site loops (Frank *et al.*, 2007).

All ThDP dependant enzymes catalyse reactions with two halves (ping-pong mechanism), the first one involving a general acid-base reaction, in which the

[1] The only exception is tartronate-semialdehyde synthase (E.C. 4.1.1.47), where the aliphatic residues surrounding the cofactor act to lower the dielectric constant of the active site, leading to activation of the cofactor by intramolecular proton rearrangement (Kaplun *et al.*, 2008; Shaanan & Chipman, 2009)

carbanion attacks an organic atom. This results in the cleavage of a bond, the attachment of the ThDP in its imino form to the substrate, and the release of the first product. The second half reaction involves a nucleophilic attack of the enamine intermediate on the second substrate and results in ligation (product). It is more diverse than the first half and depends on the substrates of the reaction (Frank *et al.*, 2007).

Catalysis follows ping-pong-mechanism kinetics, which is structurally realised by a dimer with two active sites in opposite activation states. A proton wire (over 20Å) between the two active sites passes a proton from one site to the other, toggling ThDP's activation states. The first half of ThDP-dependant enzyme reactions usually comprises a nucleophilic attack on a substrate carbonyl and general base catalysis, while the second half is specialised for the different ThDP-dependant enzymes (Frank *et al.*, 2007).

The mechanism of ThDP-dependant oxoglutarate dehydrogenase (E.C. 1.2.4.2) has two alternative routes, one of which constantly produces reactive oxygen species (ROS) and only one reaction step involving a radical, whilst the other route comprises three steps with radicals. Experiments revealed the presence of a delocalised π-radical on the enamine-thiazolium intermediate (Frank *et al.*, 2008).

Thiamine diphosphate metabolism is shared by all forms of life (Frank *et al.*, 2008). ThDP is involved in energy production, sugar metabolism and other essential pathways (Leeper & Smith, 2007). The deficiency disease of ThDP is called "beri-beri" (Leeper & Smith, 2007). In the CoFactor database, there are 30 ThDP-dependant enzymes, ten of which have a known mechanism and a MACiE entry. The 30 enzymes are mostly in one of the three classes oxidoreductases, transferases and lyases, although one hydrolase reaction (3,5/4-trihydroxycyclohexa-1,2-dione hydrolase, preliminary E.C. 3.7.1.n2) has been suggested to be added to the enzyme classification (Yoshida *et al.*, 2008).

3.2.2 Flavin adenine dinucleotide

Flavin adenine dinucleotide (FAD) is a prosthetic group in enzymes. In human metabolism, FAD is partly biosynthesised from the water soluble vitamin B_2 (riboflavin). The cofactor's molecular function comprises both one and two electron transfer reactions (Joosten & vanBerkel, 2007) and radical reactions (Buckel & Golding, 2006a).

FAD is usually non-covalently bound to the apoprotein, but can also be bound covalently to the protein[1] at one or two positions (Joosten & vanBerkel, 2007). Flavins can build stable semiquinone radicals under anaerobic conditions (Buckel & Golding, 2006a). Vitamin B_2 is the universal precursor of all flavo-cofactors (Fischer & Bacher, 2005).

The flavin biosynthetic pathway is found in plants and microorganisms (Scott *et al.*, 2007). The flavin-dependant enzyme glycine oxidase is involved in thiamine biosynthesis.

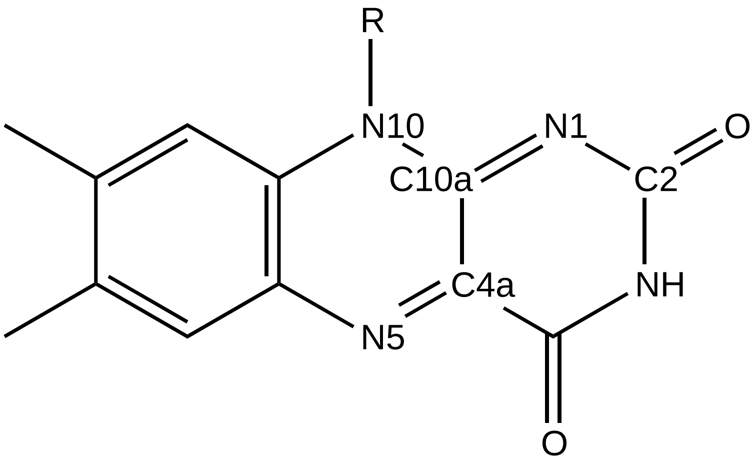


Figure 3.9: Relevant atom names in the flavin portion.

Flavin cofactors have several different mechanisms, which can be combined in various ways. It is worth noting that FAD is always regenerated in the end of the enzyme reaction. In all full mechanisms in MACiE, a net hydride transfer from or to the N5 atom (see figure 3.9) occurs during the reaction.The mechanism schemes shown on the mechanism page of the CoFactor database therefore illustrate partial reactions, rather than the overall chemical transformation. Partial mechanisms include:

1. Radical mechanism: After N5 acquires a hydride from a donor, C4a and C10a formally share a double bond. An electron acceptor then takes one of

[1] Covalent attachment may affect the redox potential of the cofactor.

the electrons from that double bonds away whereas the second one formally builds a radical at C4a. This flavin radical can either be used to channel (MACiE mechanism M0068) or temporarily store electrons (M0130) or to react with another radical (M0103). Sometimes, the unpaired electron also stays with N10 and abstracts a proton from a donor on arrival of the second single electron (M0139).

2. O-mechanism (hydride transfer): A donor provides a hydride that is attached to N5. This induces concerted double bond rearrangement in which the oxygen atom of the carbonyl group at C2 acts as an electron sink. The free electron pair of this oxygen atom then abstracts a proton from a donor. The reverse reaction then takes place such that the hydride at N5 can be transferred to the substrate (M0003).

3. O-mechanism (attachment at C4a): The flavin is again activated by a hydride transfer as described above. The free electron pair at the oxygen attached to C2 returns to C4a and attacks an electrophile which forms a bond to C4a. The former electrophile then abstracts the proton from N5. The electrophile can be a reactant (M0110) or an amino acid side chain (M0006). In the former case, the intermediate form might be a FAD radical and a reactant radical, before attachment occurs.

4. N-mechanism: In this case, the concerted double bond rearrangement stops at N1, not at the oxygen attached to C2 (M0020).

In the CoFactor database, there are 232 enzyme reactions that use FAD, 216 of which (93%) are oxidoreductases. All other classes of enzymes are found in the FAD-dependant enzymes, with the exception of ligases. The species encoding these enzymes in their genomes span all three domains of life (bacteria, archaea and eukarya).

3.2.3 Flavin mononucleotide

The Flavin mononucleotide (FMN) is mostly a prosthetic group but may also act as a coenzyme. An example mechanism, in which it acts as a coenzyme is the mechanism of alkanal monooxygenase (MACiE entry M0132). Like FAD, it is biosynthesised from vitamin B_2.

FMN is a substructure of FAD. FAD consists of the two nucleotides FMN and adenosine monophosphate, which are covalently bound through a phosphodiester bond. The flavin portion is the catalytically active portion of the cofactor. Hence, the FMN mechanisms are, as expected, similar to the FAD mechanisms.

In the CoFactor database, there are 46 enzyme reactions that use FMN, 38 of which (83%) are oxidoreductases. The species encoding these enzymes in their genomes span all three domains of life (bacteria, archaea and eukarya).

3.2.4 Nicotinamide-adenine dinucleotide

Nicotinamide-adenine dinucleotide (NAD(P)) is a collective term for two molecules performing the same function. NAD is chemically a regular dinucleotide, whereas NADP has an additional phosphate group bound at the O2' atom of the ribose ring of the adenosine monophosphate portion (see table 1.1). NAD(P) is mostly acting as a coenzyme, but has also been observed as a prosthetic group (Duine, 2001). Its main catalytic function is to assist in hydride transfers (Meijers *et al.*, 2001). Therefore the cofactor exists in two states: NAD(P)H/H^+ with, and NAD(P)$^+$ without the hydride that the cofactor shuttles (see figure 3.10).

The nicotinamide ring (pyridine ring) is planar in the oxidised form but adopts a distorted boat conformation in the reduced form (Meijers & Cedergren-Zeppezauer, 2009). This is not always visible in the PDB structures because some of the reduced NADs have been subjected to a refinement algorithm that uses standard planar restraints on the cofactor (Meijers & Cedergren-Zeppezauer, 2009). The double bond between C5 and C6 (see figure 3.10) is weakened upon adduct formation (Meijers *et al.*, 2001). The hydride transfer mechanism involving NAD(P) cofactors is accompanied by the transfer of a proton from another donor. The coupling between hydride and proton transfer has to be well orchestrated to prevent the hydride ion and proton forming a hydrogen molecule (Meijers & Cedergren-Zeppezauer, 2009).

The hydride transfer in NAD(P) mechanisms is initiated by the free electron pair of the N-atom. It is always combined with the elimination and addition of

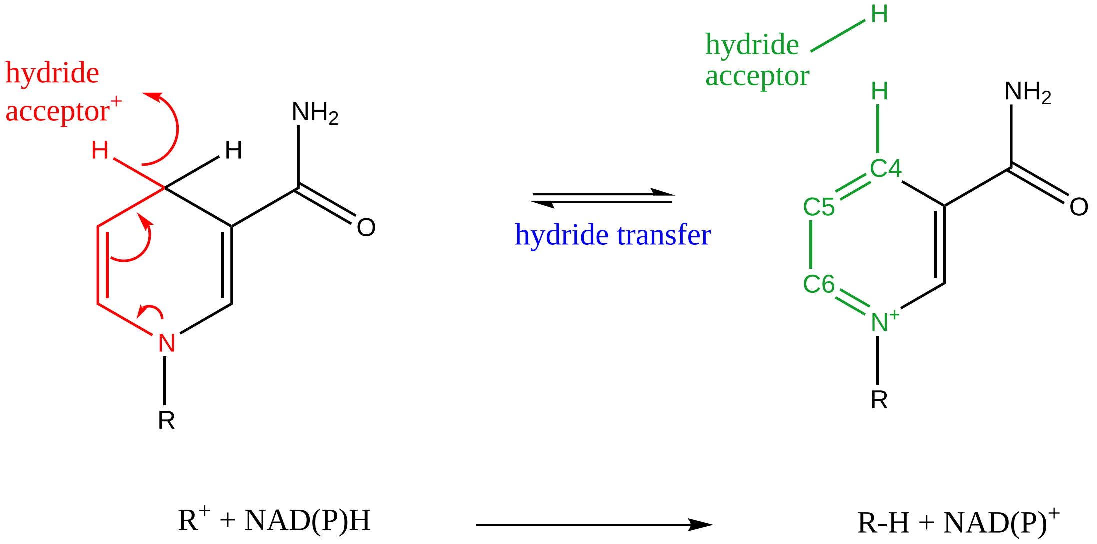


Figure 3.10: Hydride transfer mechanism of NAD(P).

the hydride. In liver alcohol dehydrogenase (E.C. 1.1.1.1), an OH-ion (ligand of a Zinc atom) activates the C4 atom of the nicotinamide ring. Therefore NAD is indirectly linked to a catalytic metal ion. The presence of a water molecule in the vicinity of nicotinamide has been found in several other protein structures (Meijers & Cedergren-Zeppezauer, 2009).

Biosynthesis of NAD(P) can be separated into two major parts: aspartic acid to nicotinic acid mononucleotide, and then to NAD(P). The biosynthetic enzymes are possible antibiotic drug targets, as human metabolism relies on the dietary uptake of one of NAD(P)'s components, vitamin B_3 (nicotinic acid) . Biosynthesis may occur through de novo synthesis or salvage pathways and they differ between pro- and eukaryotes (Rizzi & Schindelin, 2002).

NAD(P) is involved in DNA repair, calcium-dependant signalling pathways and lifespan extension in yeast (Rizzi & Schindelin, 2002). It is the most abundant electron carrier in cell metabolism (Meijers *et al.*, 2001). Further, NAD(P) is an essential cofactor for both energy metabolism and signal transduction (Rizzi & Schindelin, 2002).

The CoFactor database contains 738 NAD(P)-dependant enzyme reactions, 702 of which (95%) are oxidoreductases. The species encoding these enzymes in their genomes span all three domains of life.

3.2.5 Phosphopantetheine

Phosphopantetheine (PNS) is a prosthetic group in enzymes. In human metabolism, its biosynthesis depends on the dietary intake of vitamin B_5 (pantothenic acid). The function of this prosthetic group is to bind the initial substrate and then to function as an agile arm to bring the (growing) substrate from one active site of an multi-domain complex to the next (Sieber & Marahiel, 2005), e.g. in the fatty-acid synthase, the non-ribosomal peptide synthases and the polyketide synthases.

PNS consists of three portions connected by amide-like bonds: pantothenic acid, β-alanine and cysteamine (Koolmann & Rohm, 2003, 106). The thioester linkage between the thiol group and the acyl group being carried, provides a good leaving group for the formation of C-C bonds (Rebeille *et al.*, 2007). All crystal structures of enzymes involved in the biosynthesis of pantothenate are known (Holliday *et al.*, 2007b).

The prosthetic group is required for the biosynthesis of antibiotics, toxins and pigments (Rebeille *et al.*, 2007). It is further involved in central respiration, fatty acid synthesis and oxidation, and biosynthesis of isoprenoids (Holliday *et al.*, 2007b).

Although pantothenic acid is a vitamin for humans, there is no deficiency disease currently known in the general human population (Rebeille *et al.*, 2007). Artificial introduction of dietary deficiencies in animals have the following syndromes: depression, sleep disturbance, personality changes, cardiac instabilities and dermatitis (Rebeille *et al.*, 2007).

There are eight PNS-dependant enzyme reactions in the CoFactor database, four of which are transferases, two are oxidoreductases and one is a hydrolase and a lyase, respectively. The biological species that are annotated in SwissProt as encoding these enzyme reactions include bacteria and eukarya, but not archaea.

3.2.6 Coenzyme A

Coenzyme A (CoA) is a coenzyme and, like the phosphopantetheine prosthetic group, its biosynthesis depends on vitamin B_5. CoA may be seen as the coenzyme version of phosphopantetheine. CoA is composed of a phosphopantothenic acid portion and an adenosine mononucleotide, connected by a phosphodiester bond.

CoA's function in the cell is to solubilise hydrophobic acyl groups, *i.e.* to transport them from one enzyme to the next. Its function as a cofactor is the transfer of this acyl group. The cofactor is involved in many enzyme reactions including the citric acid cycle and also in secondary metabolism (e.g. flavonoid biosynthesis, Rebeille *et al.*, 2007).

In the bacterium *Escherichia coli*, 4% of the known enzymes have been shown to require CoA (Rebeille *et al.*, 2007). Further, evidence was found that CoA was involved in the earliest metabolic systems (Holliday *et al.*, 2007b).

The CoFactor database lists 234 enzyme reactions that use CoA, 167 of which are transferases (71%). These enzymes occur in species of all three domains of life.

3.2.7 Pyridoxal 5'-phosphate

Pyridoxal 5'-phosphate (PLP) acts as both a coenzyme and a prosthetic group. It has been described as arguably the most versatile organic cofactor (Percudani & Peracchi, 2003). In human metabolism, the cofactor's biosynthesis depends on vitamin B_6 (pyridoxal). In amino acid metabolism, PLP catalyses transamination, decarboxylation, racemisation, aldol condensation, α,β-elimination and β,γ-elimination of amino acids, and amine oxidation (Soda *et al.*, 2001). PLP covalently binds the substrate and functions as an electrophilic catalyst (Percudani & Peracchi, 2003).

Dunathan has published a study in 1966 showing that the stereo chemistry at the substrate's C2 atom in the amino acid-PLP-complex determines which bond is broken and thus which reaction occurs. Depending on which C2 bond is in a position to be activated by the PLP π-system, the cofactor either catalyses α,β-elimination, β,γ-elimination or β-decarboxylation if the C2-H bond is

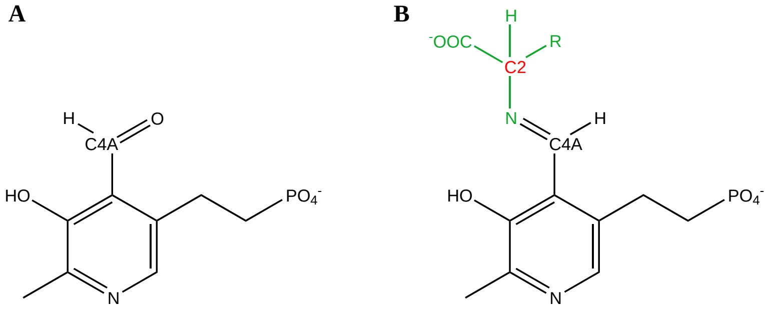


Figure 3.11: A: The PLP cofactor. B: The PLP cofactor bound to an amino acid substrate (green). The substrate's C2 atom is highlighted in red.

activated; the loss of the amino acid side chain if the C2-R bond is activated; or α-decarboxylation if the C2-COO$^-$ bond is activated (see figure 3.11B).

PLP-dependant transamination enzymes, may catalyse the hydrogen transfer from the C4A atom of the cofactor to the C2 atom of the substrate either on the *re*- or *si*- face of the planar imine intermediate (Soda *et al.*, 2001). Based on protein fold, these enzymes may be classified into four groups, which perform the following mechanisms (Soda *et al.*, 2001):

1. α- and γ-families (*si*-specific)
2. β-family (*si*-specific)
3. both specificities
4. *re*-specific

Soda *et al.* (2001) hypothesise that PLP enzymes (groups 1, 2 and 3) have mainly evolved divergently from a common ancestor, but that the *re*-face catalysing enzymes may have evolved convergently. The latter conclusion is backed up by their fold analysis.

The PLP cofactor may further function in combination with other cofactors. The enzyme 7,8-diaminopelargonic acid synthase (E.C. 2.6.1.62), for example, uses PLP as a cofactor. In the *Escherichia coli* enzyme, PLP is covalently bound to a lysine residue, while SAM, which may also act as a cofactor (see section 3.2.18), uniquely acts as an amino donor. This enzyme is the second step in biotin (see section 3.2.9) biosynthesis (Webb *et al.*, 2007). Many PLP enzymes are catalytically promiscuous (*i.e.* the same enzyme catalyses more than one reaction, Percudani & Peracchi, 2003).

PLP biosynthesis occurs on two alternative pathways in different species (Scott *et al.*, 2007).

PLP is mainly involved in metabolism of amino compounds, but also co-catalyses the reactions of glycogen phosphorylases (E.C 2.4.1.1, (Percudani & Peracchi, 2003), Percudani & Peracchi, 2003). PLP occurs in all organisms. 1.5% of all genes in free-living prokaryotes encode PLP-dependant enzymes. This indicates that PLP is mainly involved in basic metabolism (Percudani & Peracchi, 2003).

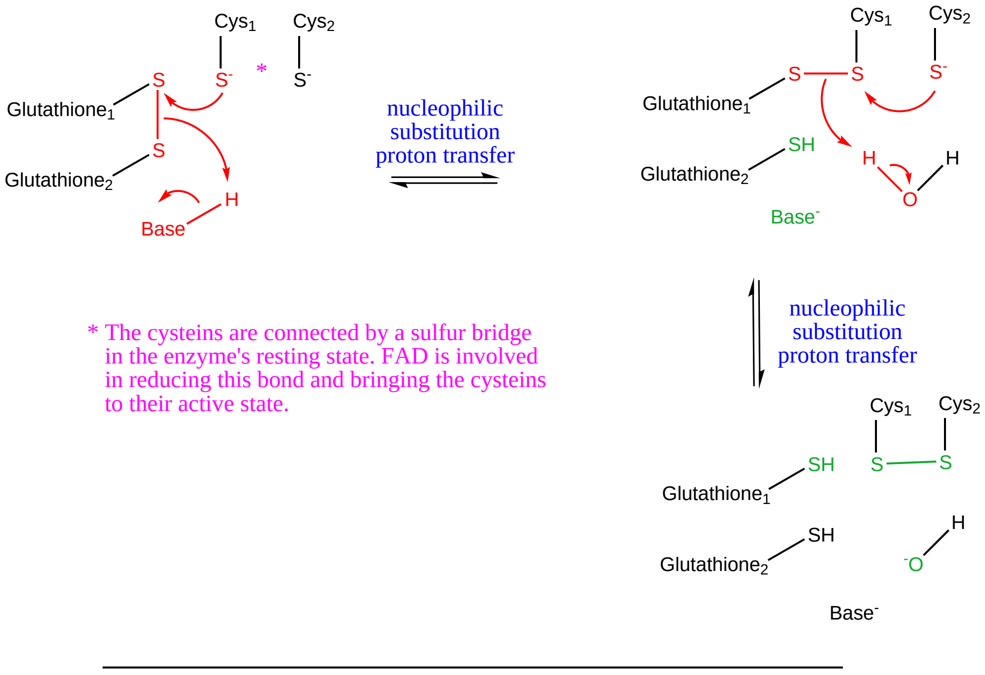
It has been shown for some enzymes that the decreased catalytic activity remains with the cofactor, not the enzyme, when cofactor and enzyme are separated (Gramatikova *et al.*, 2002).

The CoFactor database lists 139 PLP-dependant enzyme reactions, 73 of which (53%) are transferases and 52 of which (37%) are lyases. These enzymes are encoded in the genomes of species from all three domains of life.

3.2.8 Glutathione

Glutathione (GSH) is a coenzyme that is biosynthesised in human metabolism. The molecule is a tripeptide comprising glutamate, cysteine and glycine (Aoyama *et al.*, 2008), where glutamate and cysteine are linked through a gamma-glutamyl amide linkage (Dalton *et al.*, 2004). GSH functions as an antioxidant in the cytoplasm in general (Van Petegem *et al.*, 2007) and in neurons (Aoyama *et al.*, 2008) in particular. Its chemical function is to act as an electron donor. The cofactor also serves as a means of cysteine storage in the cell. This is necessary, because pure cysteine is very reactive and neurotoxic and may lead to free radical generation. It is estimated that half of liver GSH serves as a cysteine reservoir (Aoyama *et al.*, 2008). Glutathione counteracts oxidative stress, both when bound to enzymes as well as in solution in the cell. Non-enzymatically, GSH reacts with nitrous oxide, hydroxyl radicals and $ONOO^-$. There is no known enzymatic defence against hydroxyl radicals (Aoyama *et al.*, 2008). It may further act as a redox buffer and take part in the detoxification of heavy metals (Dalton *et al.*, 2004). The cofactor may not only be an antioxidant, but might also have an oxidising effect: it acts as an electron donor to metal cations where O_2 is the electron acceptor. This process produces superoxide anions and eventually H_2O_2 (Pompella *et al.*, 2003).

In addition to the cofactor's role as a coenzyme, GSH is a neuromodulator in the central nervous system (Aoyama *et al.*, 2008). It is further involved in cell fate decisions such as proliferation and apoptosis (Dalton *et al.*, 2004). The cysteine residues of proteins may be glutathionylated, which may have regulatory effects that may also influence development when applied in transcription factors. Additionally, it serves to protect protein cysteine residues (Pompella *et al.*, 2003).



* The cysteins are connected by a sulfur bridge in the enzyme's resting state. FAD is involved in reducing this bond and bringing the cysteins to their active state.

$2\ GSSG + 2H^+ + 2e^- \ \text{<==>}\ 2\ GSH$

Figure 3.12: The recycling reaction of glutathione: from the oxidised form GSSG to the reduced forms GSH.

Generally, GSH is a nucleophile and a reductant (Pompella *et al.*, 2003). Two molecules GSH may be connected *via* a disulfide bond (GSSG), which is the oxidised form of glutathione. When this molecule is cleaved, two molecules GSH (reduced, antioxidant function) are obtained (see figure 3.12). Glutathione amide disulfide (GASSAG) may act in an analogous manner (Van Petegem *et al.*, 2007). The oxidised/reduced ratio in the cell is 1:100 under normal conditions or 1:49 under stress conditions (Dalton *et al.*, 2004).

Glutathione biosynthesis occurs only in the cytoplasm, but the molecule is located also in mitochondria. Biosynthesis requires two steps, which are both catalysed by ATP-dependant ligases: γ-glutamylcysteine synthase and glutathione synthase. The cofactor regulates its own biosynthesis via feedback inhibition (Aoyama *et al.*, 2008). Glutaminase (E.C. 3.5.1.2) transforms glutamine to glutamate for GSH biosynthesis, and the trans-sulfuration pathway provides cysteine from methionine (Aoyama *et al.*, 2008).

GSH participates in nucleotide metabolism and the formation of lipid second messengers. It is a neurotransmitter and regulates nitric oxide homeostasis (Dalton *et al.*, 2004). GSH executes its antioxidative function often in the brain, which produces ROS due to its high usage of oxygen (Van Petegem *et al.*, 2007).

Humans biosynthesise GSH. There are genetic defects and polymorphisms that result in mild or severe modifications of the GSH level, and respective symptoms, yet these are rather rare (Dalton *et al.*, 2004). GSH deficiency in the brain is thought to be connected with Parkinson's disease, Alzheimer's disease and amyotrophic lateral sclerosis, because a higher neuronal level of GSH concentration could eliminate oxidative stress, causing these age-related neurodegenerative diseases (Aoyama *et al.*, 2008). Orally administered GSH is hydrolysed by dipeptidase and if intravenously administered, it is rapidly eliminated by reaction with γ-glutamyl-transferase (E.C. 2.3.2.2, half life 7 min, Aoyama *et al.*, 2008).

In addition to its role in neurodegenerative diseases, GSH is involved in cancer (S-Glutathionylation of p53), cystic fibrosis, HIV (S-Glutathionylation of HIV-1 protease) and ageing (Dalton *et al.*, 2004). In the mouse model, genetically GSH deficient mice have high DNA damage increasing with age.

The CoFactor database contains GSH-dependant enzyme reactions from bacteria and eukarya, but not from archaea (possibly due to lack of annotation). 18 (51%) of the 35 GSH-using enzyme reactions in CoFactor are oxidoreductases.

3.2.9 Biotin

The prosthetic group biotin (BTN) is a vitamin for humans (vitamin B_7). The molecule is always covalently bound to its partner enzyme. Its main function as a cofactor is to transfer CO_2 from one active site of an enzyme complex to another or to act as a CO_2-carrier between bicarbonate and the acceptor substrate. It may also transfer C_2-units. BTN is essential for fatty acid synthesis (Rebeille *et al.*, 2007).

The BTN requirement in the cell is very low, so the only nutritional deficiency is found in patients with a diet rich in raw egg-white. The avidin from raw egg white has a high affinity to biotin and thus depletes the BTN pool. Other related diseases are mainly genetic disorders, resulting from a loss of function of biotinidase (E.C. 3.5.1.12, the recycling enzyme) or holocarboxylase synthetases (the biotin inserting enzymes, Rebeille *et al.*, 2007).

The cofactor is biosynthesised in plants and bacteria (Rebeille *et al.*, 2007). In the second step of the biosynthesis process is the aforementioned (3.2.7) enzyme 7,8-diaminopelargonic acid synthase, which utilises PLP and SAM (see section 3.2.18), but SAM as an amino donor (Scott *et al.*, 2007). The sulfur atom of biotin is donated by an iron sulfur cluster (Leeper & Smith, 2007; Webb *et al.*, 2007).

Part of the BTN that is required by human metabolism is produced by the intestinal flora. Further, the enzyme biotinidase recycles the cofactor from enzymes gained from food sources. A biotin deficiency causes multi-carboxylase deficiency, which may result in organic acideria, a metabolic disorder that disrupts amino acid metabolism. Other related diseases include neurological disorders and developmental delay (Rebeille *et al.*, 2007).

Biotin occurs in all three domains of life. 7 (64%) out of the 11 biotin-dependant enzyme reactions in the CoFactor database are ligases. The others are lyases and one transferase.

3.2.10 Tetrahydrofolic acid

In human metabolism, the coenzyme tetrahydrofolic acid (THF) is the essential nutrient vitamin B_9. The molecule is composed of a pterin ring, a p-aminobenzoic acid and a polyglutamate chain (Leeper & Smith, 2007).

The cofactor performs a crucial function in C_1-metabolism: it transports and donates C_1 groups (*i.e.* one-carbon moieties). This includes mainly methyl groups from serine hydroxymethyltransferase reaction (E.C. 2.1.2.1), but in general the C_1 moiety may be reduced or oxidised (methyl, methylene, formyl or methenyl). Thus, the folate pool is a mixture of molecules, differing in the oxidation state of the pterin ring (di- or tetrahydro), the oxidation state of the C_1-unit carried, and in the length of the glutamate chain. Only the tetrahydro state can transport C_1 units (Leeper & Smith, 2007). The folate molecule is similar to a number of other pterin-based cofactors, like molybdopterin (see section 3.2.15) or flavins.

Folic acid cofactors are involved in purine and pyrimidine biosynthesis. Thus, the synthesis of RNA, DNA and the cofactors NAD(P), FAD, CoA and SAM depend on them. The C_1-units from the folate cofactors become C2 and C8 of the purine ring (see figure 3.13). Folate derivatives also act as chromophores in DNA-repair (Leeper & Smith, 2007).

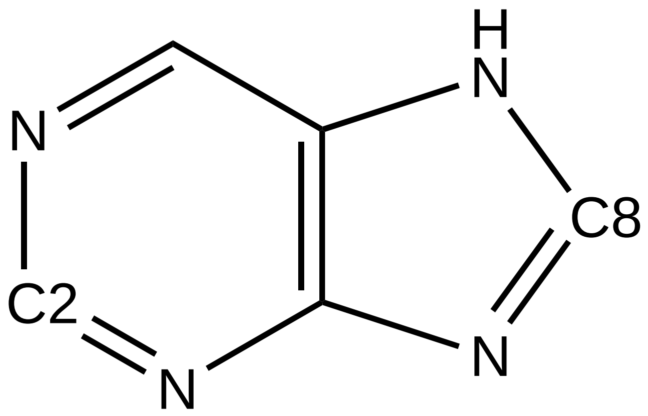


Figure 3.13: C2 and C8 in a purine ring: the one-carbon-moieties delivered by the THF cofactor.

Methionine synthase (E.C. 2.1.1.13) uses THF and is a crucial enzyme for the regeneration of methionine from homocysteine. Methionine is required for the cofactor SAM (see section 3.2.18), which is transformed to the toxic component homocysteine when SAM is consumed in methylation reactions. The enzyme has two isozymes, one of which depends on cobalamin (see section 3.2.11) and one of which does not (Leeper & Smith, 2007).

THF is biosynthesised in microorganisms including bacteria, in lower eukaryotes and plants, but not in animals. It is synthesised from GTP, chorismate and glutamate, and there are 11 enzymes in the biosynthetic pathway (Holliday *et al.*, 2007b). Green leafy vegetables are a good food source for folate, in contrast to roots (like carrots), storage organs (like potatoes) or most fruits. Folate content highly depends on food storage, processing and cooking (Leeper & Smith, 2007).

The CoFactor database contains five enzyme reactions that use the THF cofactor, four of which are transferases. All three domains of life have THF-dependant enzymes.

3.2.11 Adenosylcobalamin

Adenosylcobalamin (B12) is a prosthetic group, whose source for humans is dietary (vitamin B_{12}, cobalamin) . The cofactor assists enzymes with the catalysis of molecular rearrangements, methylations and dehalogenations (Mendel *et al.*, 2007).

It is worth noting that vitamin B_{12} is a collective term for a group of similar molecules, the corrinoids. They all have the corrin ring and the central cobalt atom in common, while they differ in the two axial ligands of the central metal atom. The upper (β) axial ligand is either an adenosyl- or a methyl-moiety, while the lower (α) axial ligand is attached to the corrin ring. In animals, this lower ligand is 5,6-dimethylbenzimidazole, but adenine has also been observed (cofactor also know as pseudo-B_{12}). A summary and further references can be found in the publication of Taga & Walker (2008).

The cobalt atom is coordinated by four nitrogen atoms from the corrin ring, which is one C-atom less than in the other corrin cofactors (hemes and factor F430, see sections 3.2.20 and 3.2.19). The shortening allows for a tighter coordination of the cobalt atom. The axial ligands are the endogenous ligands, 5,6-dimethylbenzimidazole and an adenosyl portion. The molecule is highly methylated. This encourages the molecule to adopt the appropriate conjugation state and prevents prototrophic rearrangement and oxidation (Holliday *et al.*, 2007b).

B12-dependant enzymes occur predominantly in anaerobic organisms since the molecule is sensitive to oxygen. An exception is the enzyme methylmalonyl-CoA mutase (E.C. 5.4.99.2), which is involved in propionate formation (in propionibacteria) and in proprionate oxidation (in syntrophic bacteria and animals). In anaerobic organisms, eliminases and mutases may be B12-dependant (Buckel & Golding, 2006a).

In the enzymes glycerol dehydratase (E.C. 4.2.1.30) and lysine-5,6-aminomutase (E.C. 5.4.3.3 and 5.4.3.4), B12 becomes inactivated during the turnover and is recycled by a chaperon-like reactivation factor (ATP-dependant, Buckel & Golding, 2006a).

B12-dependant enzymes can be classified into three classes: isomerases (A), methyltransferases (B) and reductive dehalogenases (C). Group A mainly catalyses 1,2-rearrangements, often in anaerobic fermentative processes (Leeper & Smith, 2007).

In all B12-dependant enzyme reactions, the function of B12 is to generate the reactive adenosyl radical by homolytic cleavage of the Co-C bond. This radical then reacts with the substrate and is eventually quenched by the re-formation of the Co-C bond.

It is not clear how old B12 is in evolutionary terms. On one hand, it has a nucleotide loop which could be a hint that it was involved in ribozymes in a potential RNA world (Benner *et al.*, 1989). Furthermore it is involved in nucleotide reduction, which is a very basic process in nature (Dickman, 1977). On the other hand, the enzyme ribonucleotide reductase (E.C. 1.17.4.2) exists in three different variations, one B12-dependant, the other two using SAM and molecular oxygen (Nordlund & Reichard, 2006). The SAM-dependant enzyme seems to be the oldest (Holliday *et al.*, 2007b).

B12 is only synthesised in prokaryotic organisms. Higher plants do not require it, because they have an alternative form of methionine synthase. Some bacteria in the lower intestine produce a form of B12 that cannot be taken up by the human metabolism, hence most B12 has to be taken up with the diet: meat, dairy products and eggs are good sources of B12 (Leeper & Smith, 2007). Vegetarians can retrieve it e.g. from algae in sushi.

Cobalt is a trace element that is needed in very small quantities, mainly for the biosynthesis of vitamin B_{12} (or its uptake) (Mendel *et al.*, 2007). B12 deficiency results in pernicious anaemia. In humans it affects the two enzymes that depend on it: methionine synthase and methylmalonyl-CoA mutase.

Approximately 30 enzymes are in involved in the biosynthesis of vitamin B_{12}. 25 of those are specific to B12 biosynthesis, the other ones are common between B12 and the other corrin cofactors (Mendel *et al.*, 2007). There are at least two routes for biosynthesis: aerobic and anaerobic (Holliday *et al.*, 2007b). In contrast to the same phenomenon in e.g. heme biosynthesis, here the two pathways have unique intermediates. The difference between the two routes is the method of ring contraction and the moment of metal insertion (Mendel *et al.*, 2007).

The cobalt uptake proteins probably have a common origin with nickel uptake proteins. Many potential cobalt transport systems are under B12-riboswitch control (Holliday *et al.*, 2007b).

The CoFactor database contains 15 B12-dependant enzyme reactions, 7 of which are isomerases. Species in all three domains of life encode enzymes with these E.C. numbers in their genomes.

3.2.12 Ascorbic acid

The coenzyme ascorbic acid (ASC) is also known as the water-soluble vitamin C. The molecule is important in the cell's antioxidant defence, particularly *via* the ascorbate/glutathione cycle (Ishikawa & Shigeoka, 2008). Ascorbic acid is the major water-soluble antioxidant found in body fluids of mammals. It counteracts free or lipid-peroxidation-initiating radicals and it regenerates other antioxidants like Vitamin E. Extracellular ascorbate in its reduced form is recycled by coenzyme Q (see section 3.2.14), which takes place at the plasma membrane (Ishikawa & Shigeoka, 2008). Vitamin C is involved in scavenging of superoxide, hydroxyl radicals and singlet oxygen, directly as well as in enzymes (Smirnoff, 2000).

In mammals, the eye contains one of the highest concentrations of ASC, to act as an antioxidant in the light-absorbing (and thus ROS-producing) processes. The cofactor is, apart from some inorganic ions, the most abundant molecule in chloroplasts (Smirnoff, 2000).

ASC exists in three forms: fully oxidised (dehydroascorbate), as a radical after a one-electron transfer (semidehydroascorbate) and fully reduced (ascorbate). This is mirrored in the mechanism, which is a two step process involving two intermediates: firstly the one-electron intermediate ascorbate free radical and secondly the fully oxidised dehydroascorbate (Arroyo *et al.*, 2004). In the mechanism, two reduced ASC molecules are used and thus transformed to two dehydroascorbate radicals, which then react with each other to produce one fully oxidised dehydroascorbate and one fully reduced ASC molecule (disproportionation reaction, Smirnoff, 2000).

Scurvy is the human vitamin C deficiency disease. ASC-deficient plants are hypersensitive to oxidative stress (Smirnoff, 2000).

In the majority of E.C. 1.14.11.- sub-subclass of enzymes, the IntEnz database lists ascorbate as a cofactor. Muller *et al.* (2004) have reviewed that the role of ascorbate is still unclear, as it is required stoichiometrically by some of the enzymes and has been shown to accelerate the reaction in all cases where the kinetic effect has been characterised. The activity of some of these enzymes does not require ASC but is strongly stimulated by it. The authors further summarise that it has been proposed to play a part in the prevention of oxidative self-inactivation, but the mechanism has yet to be elucidated.

19 out of the 20 ASC-using enzymes in the CoFactor database are oxidoreductases. These E.C. numbers are annotated in SwissProt to occur in bacteria and eukarya, but no archaeal protein sequences are currently annotated with any of these E.C. numbers.

3.2.13 Menaquinone

The coenzyme menaquinone (MQ) is taken up into human cells and known as vitamin K_2.

According to Oldenburg *et al.* (2008), “vitamin K is a collective term for lipid-like naphthoquinone derivatives synthesised only in eubacteria and plants and functioning as electron carriers in energy transduction pathways and as free radical scavengers maintaining intracellular redox homeostasis.” Vitamin K is phylloquinone (K_1) or menaquinone (K_2). They further summarise that "phylloquinone acts as the electron transfer cofactor A1 of photosystem I in plants, while menaquinone plays an essential role in several microbial electron transport systems" (Oldenburg *et al.*, 2008).

MQ shuttles reducing equivalents as electrons between two enzymes: vitamin K epoxide reductase complex subunit 1 and γ-glutamyl carboxylase (E.C. 4.1.1.90). In the latter enzyme, MQ catalyses the formation of γ-glutamyl carboxylated residues from specific glutamate residues in certain proteins. These modified residues confer calcium-binding properties (Booth & Al Rajabi, 2008). Generally, MQ is involved in energy transduction pathways (as an electron carrier) and in blood coagulation. It further functions as as free radical scavenger (Oldenburg *et al.*, 2008).

The cofactors ThDP (Oldenburg *et al.*, 2008) and SAM (Meganathan, 2001) are involved in vitamin K_2-biosynthesis.

The CoFactor database contains only one enzyme reaction for MQ, which is the aforementioned γ-glutamyl carboxylase. This reaction is not annotated in SwissProt at all, but it is known that humans have this enzyme.

3.2.14 Ubiquinone

The coenzyme ubiquinone is also known as coenzyme Q (CoQ). The molecule is fully biosynthesised in human cells. CoQ has two major functions in the cell: to transport electrons in the respiratory chain and to act as a lipid soluble antioxidant (Tran & Clarke, 2007). This molecule is one out of only four fat-soluble antioxidants in nature, the others being carotinoids, oestrogens and tocopherols (vitamin E). CoQ inhibits protein oxidation of basic amino acids and SH-groups by breaking the radical chain reaction. It is also involved in inhibiting lipid and DNA oxidation. The cofactor occurs in all human tissues. Generally, CoQ is involved in energy metabolism (respiratory electron transport chain) and in oxidative stress response (Tran & Clarke, 2007).

Like MQ, CoQ has three states: quinol (reduced), semi-quinone (radical), and quinone (oxidised). The cofactor may transfer up to two electrons.

Genetic deficiencies, ageing or damage of the CoQ-biosynthetic system by diseases or toxic injury are possible reasons for low CoQ levels. Healthy tissues synthesise enough CoQ such that uptake is unnecessary. The uptake capacity is therefore very limited. In the case of CoQ deficiency, biosynthesis needs to be stimulated in addition to the uptake system. Candidate drug molecules are metabolic intermediates (Bentinger *et al.*, 2007).

All five CoQ-using enzyme reactions in the CoFactor database are oxidoreductases. The cofactor is used in all three domains of life.

3.2.15 Molybdopterin

Molybdopterin (MTE) is a prosthetic group in enzymes, which is biosynthesised in human metabolism. The cofactor coordinates a molybdenum atom using the two sulfur atoms of the tricyclic pterin portion. During catalysis, the ring system participates in electron transfer. This allows for modulation of the reduction potential (Mendel *et al.*, 2007; Rebeille *et al.*, 2007). Pterin has several reduction states and conformations and the Mo atom shuttles between oxidation states +IV and +VI during catalysis (Rebeille *et al.*, 2007). This may be important for the channelling of electrons to other prosthetic groups.

The tricyclic pterin portion often occurs connected to a cytosine- or guanine-nucleotide, where the two molecules may be connected *via* a phosphodiester bond in eubacteria (Buckel & Golding, 2006b). Two molecules of the molybdenum cofactor (MoCo) are required to provide four sulfur atoms for molybdenum-coordination. The metal molybdenum is only biologically active when it is complexed with its cofactor (except for the bacterial nitrogenase, Mendel *et al.*, 2007). The advantage of using molybdenum is the possibility to carry out many-electron reactions at a low potential (Holliday *et al.*, 2007b). Since the availability of molybdenum in the pre-biotic world is not confirmed, it has been hypothesised that tungsten or vanadium used to carry out similar tasks (Holliday *et al.*, 2007b).

In terms of pathways, MoCo is involved in the nitrogen cycle, the sulfur cycle and the carbon cycle (Rebeille *et al.*, 2007). Enzymes using this cofactor usually catalyse oxo-transfers with subsequent two-electron oxidation or reduction of the molybdenum ion. The catalytic cycle further requires an electron transfer involving the molybdenum-centre and additional redox cofactors like iron sulfur clusters, flavins or hemes. Regarding mechanism, it is known that aldehyde dehydrogenase proceeds *via* a trihydropterin radical. Aldehyde ferrodoxin oxidoreductase (E.C. 1.2.7.5) might generate pterin radicals in the process of oxidising ferrodoxin (Wei *et al.*, 2003).

When MoCo is outside the protein, it loses the molybdenum atom and is quickly oxidised (and from thereon inactive). This oxygen-sensitivity suggests that MoCo is unlikely to be seen freely in the cell, although a carrier protein been characterised in an algae species. It is known that tungsten, an antagonist, inhibits MoCo activity, although it is the natural ligand in hyperthermophilic archaeal bacteria (Mendel *et al.*, 2007).

Insertion of the cofactor into the apoprotein is not understood in detail. Usually MoCo is inserted after all the other redox cofactors. In some MoCo enzymes, the pterin ring system may be covalently linked to a protein cysteine residue. There is also an iron containing MoCo in bacterial nitrogenase (Mendel *et al.*, 2007). There are more than 50 molybdenum-containing enzymes known, most of them bacterial (Rebeille *et al.*, 2007).

Due to homology of MoCo biosynthesis to ubiquitin biosynthesis, it is likely that ubiquitin-dependant degradation has evolved from the evolutionarily older MoCo biosynthesis pathway (Holliday *et al.*, 2007b). Copper is involved in molybdenum metabolism, possibly as a place holder for molybdenum, until it is inserted (Mendel *et al.*, 2007).

Molybdenum uptake in bacteria functions via ATP-binding cassette transporters. In higher organisms, molybdenum possibly uses the phosphate transport. *Arabidopsis thaliana* has a distinct molybdenum uptake system (Mendel *et al.*, 2007).

The structural homology between molybdopterin synthase and ThiS (an enzyme from thiamine synthesis) unambiguously demonstrate the divergent evolutionary relationship between a subset of enzymes involved in the biosynthesis of sulfur-containing cofactors. Both proteins share the same fold as ubiquitin (Rizzi & Schindelin, 2002).

Mutations in the first two biosynthesis enzymes cause severe and progressive neurological abnormalities. The disease is not treatable, because MoCo is not stable outside its enzymes. Possible future treatments include the supplementation of pyranopterin monophosphate, which is the product of the first step of biosynthesis. This therapy may only help those patients that have a mutation in this first enzyme (Rebeille *et al.*, 2007).

MoCo-using enzymes may be classified based on the homology of their active centres (Mendel *et al.*, 2007):

1. Xanthine oxidase family: hydroxylases with a monoxy-molybdenum centre, the only class found in higher organisms such as human (Rebeille *et al.*, 2007)

2. Sulfite oxidase family: oxotransferases with a dioxy-molybdenum centre

3. DMSO reductase family: two pterin molecules coordinate one molybdenum centre

4. Aldehyde ferrodoxine reductase family (Wei *et al.*, 2003)

MTE occurs in all three domains of life. All 14 MTE-dependant enzymes in the CoFactor database are oxidoreductases.

3.2.16 Biopterin

Tetrahydrobiopterin (BH4), like molybdopterin, is a cofactor that contains a pterin portion. The cofactor is biosynthesised in human cells and may act as both, coenzyme or prosthetic group. The cofactor performs one- or two-electron transfers (Wei *et al.*, 2003) and the OH-group transfer for aromatic amino acids (Thony *et al.*, 2000). It is also a cofactor of nitric oxide synthase (E.C. 1.14.13.39), which is involved in neuronal signalling. In this enzyme, BH4 is regenerated in the reaction (Wei *et al.*, 2003). It is most likely involved in catalysis as a pterin radical (Thony *et al.*, 2000).

BH4 belongs to a family of similar molecules (pterins, lumaines, alloxazines, folates and riboflavins). Their core structure is composed of two or three heterocyclic six-membered rings (Wei *et al.*, 2003). The protonation state of BH4 is thought to regulate the cofactor's binding and function in enzymes (Wei *et al.*, 2003). Through the O4 and N5 atoms, pterins may form metal chelates, but these complexes do not mimic how molybdopterin and tungstenpterin are bound (Wei *et al.*, 2003).

BH4 also reduces the highly reactive and therefore dangerous peroxynitrite (non-enzymatically). In the course of the reduction reaction, the cofactor BH4 is reduced *via* the BH3 radical to become BH2, which is later recycled by dihydrofolate reductase (Schmidt & Alp, 2007). BH4 is essential for phenylalanine metabolism (Thony *et al.*, 2000). The BH2 and BH4 states are shown in figure 3.14.

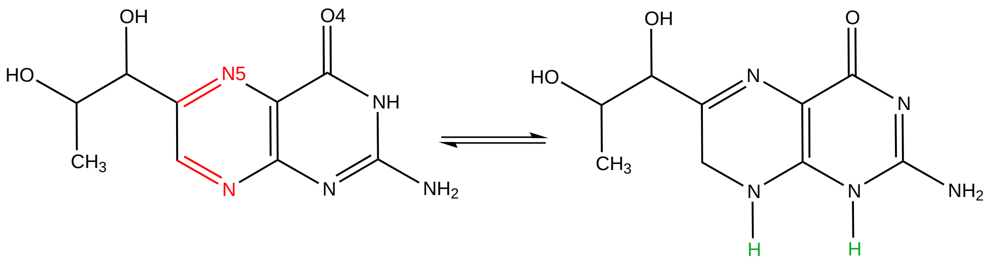


Figure 3.14: The reduced (BH2) and oxidised (BH4) state of biopterin.

Mutations in most of the enzymes cause hyperphenylalaninaemia, a form of phenylketonuria. Restricted cofactor availability has also been suggested to be involved in Alzheimer's disease, Parkinson's disease, autism, depression and dihydroxyphenylalanine-responsive dystonia (Thony *et al.*, 2000).

In phenylalanine metabolism, reduced BH4 assists in a reaction that uses molecular oxygen resulting in hydroxylation of BH4 (BH4-4α-carbinolamine). If the hydroxylated form of the cofactor accumulates, BH4-4α-carbinolamine produces harmful metabolites. Thus, the enzymatic regeneration of the cofactor is essential here. The three amino acid hydroxylases that use BH4 depend additionally on iron and O_2 (Thony *et al.*, 2000).

In the aromatic amino acid hydroxylases, biopterin functions differently than in nitric oxide synthase (Wei *et al.*, 2003):

1. The cofactor is oxidised by two electrons rather than by one

2. It participates directly in oxygen activation

3. It is released from the enzyme after each catalytic cycle (and thus acts as a coenzyme, rather than a prosthetic group)

Interestingly, no information has been documented that individuals lacking glyceryl-ether monooxygenase show any deficiency symptoms (Thony *et al.*, 2000).

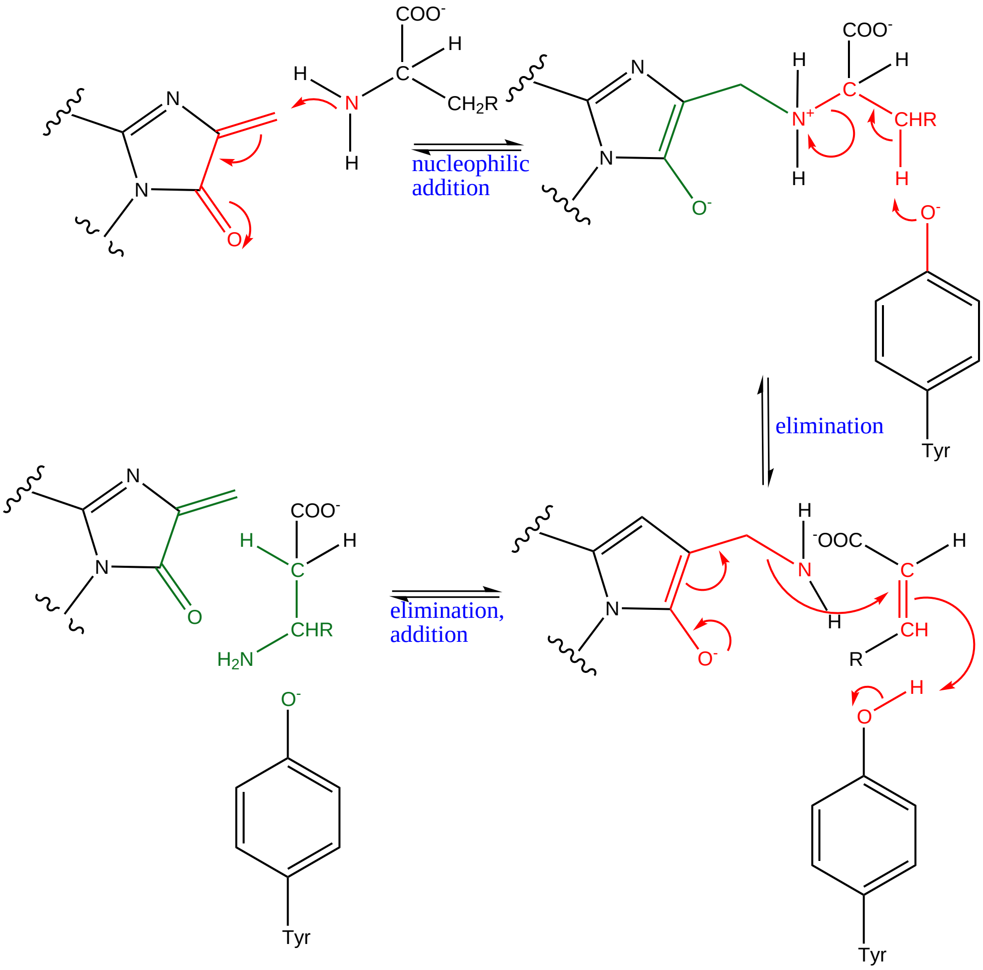
All 6 BH4-dependant enzymes in the CoFactor database are oxidoreductases and all of them have only been annotated in eukarya in SwissProt.

3.2.17 4-methylideneimidazole-5-one

The molecule 4-methylideneimidazole-5-one (MIO) is a prosthetic group in enzymes. To be precise, it is a special case of prosthetic group, namely a polypeptide-derived cofactor (see section 1.3.1). For this cofactor there are no biosynthetic enzymes required since it arises from auto-catalysis of the three amino acids alanine, serine and glycine in the enzyme sequence. The self-catalysis process is chemically analogous to chromophore formation. A phenylalanine residue is proposed to play a key role in auto-catalysis of MIO (Christianson *et al.*, 2007c).

MIO functions as a strong electrophile to abstract a relatively non-acidic hydrogen atom (Christianson *et al.*, 2007a) from the substrate. It catalyses a 1,2-shift of amine substrates (see figure 3.15), which is chemically challenging (Christianson *et al.*, 2007c). Two mechanisms have been proposed: after the removal of the relatively non-acidic β-hydrogen, the reaction pathway splits into two possibilities. In path A, the cofactor reacts with α-amino group, whereas in path B it reacts with the aromatic ring. In both cases, the next step performs a deprotonation and elimination of ammonia, which leads to conjugated olefin. Christianson *et al.* (2007a) have provided experimental evidence consistent with path A.

Tyrosine 2,3-aminomutase (E.C. 5.4.3.6, mechanism see figure 3.15) uses MIO, where other aminomutases use other cofactors instead. Lysine-2,3-aminomutase (E.C. 5.4.3.2), for instance, uses a combination of SAM (see section 3.2.18), an iron sulfur cluster and PLP. Other aminomutases use B12 and PLP. The histidine ammonia lyase family is thought to have a common catalytic pathway with tyrosine 2,3-aminomutase (Christianson *et al.*, 2007c). Further, natural product



C_{α}–amino–Substrate + MIO ⟶ C_{β}–amino–Substrate + MIO

Figure 3.15: The mechanism of tyrosine 2,3-aminomutase (see also MACiE M0245).

biosynthesis pathways for enedines, taxanes, non-ribosomal peptides contain MIO (Christianson *et al.*, 2007a).

Of the four MIO-dependant enzymes in the CoFactor database, three are lyases and one is an isomerase. These E.C. numbers have only been annotated in eukarya in SwissProt.

3.2.18 S-adenosylmethionine

In human cells, the coenzyme S-adenosylmethionine (SAM, also known as AdoMet) is biosynthesised. SAM is the "major and most commonly used methyl group donor in all biological systems" (Arroyo *et al.*, 2004). The coenzyme methylates DNA, RNA, proteins and small molecules. SAM may also catalyse radical reactions using an adenosyl-radical, in which the radical is formed by a one-electron transfer from flavodoxin or ferrodoxin (Buckel & Golding, 2006b), which then abstracts a hydrogen atom either directly from the substrate or from a glycine residue (Rebeille *et al.*, 2007). Unlike other SAM enzymes, 7,8-diaminopelargonic acid synthase (E.C. 2.3.1.47) uses SAM as an amino donor. This enzyme also uses PLP as a cofactor and is the second step in biotin biosynthesis (Holliday *et al.*, 2007b). According to the cofactor definition used in this work (see section 1.3.1), SAM is not seen as a cofactor in this reaction.

SAM contains a reactive sulfur cation, which often acts as a methyl-group donor. One product of such a reaction is S-adenosyl-homocysteine, which has a cytotoxic effect. The cofactor is further involved in methionine generation, in which case a 5'-deoxyadenosyl radical is generated (Roje, 2006). Methionine synthase exists in two isozymes. Both depend on a folate cofactor as a methyl-group donor and one (MetH) is additionally dependant on B12 (Rebeille *et al.*, 2007). In bacteria there is a SAM-riboswitch, which regulates cysteine and methionine synthesis (Cochrane & Strobel, 2008). SAM is further a co-repressor of methionine biosynthesis (Roje, 2006). The cofactor occurs in anaerobes, but also in aerobic bacteria, fungi, plants and animals (Buckel & Golding, 2006b). In plants, SAM is also a precursor of polyamines, nicotinamines, ethylene, and source of 5'-deoxyadenosyl radicals (Roje, 2006).

In the coenzyme, the methyl group is bound to a charged sulfur atom, which thermodynamically destabilises the molecule. This results in a high leaving group potential for the methyl-group (Turner *et al.*, 2000).

SAM catalyses a 1,2-shift of substrate amines (Christianson *et al.*, 2007d) and is a cofactor in lysine-2,3-aminomutase and spore photoproduct lyase (E.C. 4.1.99.14) (Buckel & Golding, 2006b). DNA methylation by various different restriction enzymes requires SAM (Arroyo *et al.*, 2004).

Glycyl-radical enzymes are activated by an irreversible hydrogen abstraction by SAM (e.g. pyruvate:formate lyase). They occur only in anaerobes, because the reaction with dioxygen would cleave the protein chain. The five known glycyl-radical enzymes are (Buckel & Golding, 2006b):

1. pyruvate:formate lyase, (closely related to 2-oxobutyrate formate lyase)
2. anaerobic or type III ribonucleotide reductase
3. benzylsuccinate synthase
4. p-hydroxyphenylacetate decarboxylase
5. B12-independant glyceroldehydratase

SAM can be reductively cleaved by an iron sulfur cluster (Klinman, 2001). The cofactor is involved in biotin synthesis (Roje, 2006) as well as in heme (HemN) and molybdopterin (MoaA) biosynthesis. All three enzymes have iron sulfur clusters for single electron transfer for SAM-activation (Buckel & Golding, 2006b).

The CoFactor database contains 161 SAM-dependant enzyme reactions, 153 of which (95%) are transferases. SAM occurs in all three domains of life.

3.2.19 Factor F430, Coenzyme M and Coenzyme B

Factor F430 (F430) is a prosthetic group, which only occurs in methanogenic bacteria and is involved in one reaction (coenzyme-B sulfoethylthiotransferase, E.C. 2.8.4.1) in methanogenesis. Its function there is to bind and later to release the methyl leaving group from coenzyme M. In the same reaction, F430 also performs the initiation and termination of a radical reaction, and coenzyme B is also involved (see figure 3.16 and MACiE entry M0156 for mechanism). Although there are two alternative mechanisms proposed, the one described here is better supported by the data available. There, the nickel attacks the methyl group of coenzyme M, which results in the methyl group transfer to F430. The mechanism proceeds via radical intermediates and finally yields free methane, and covalently connected coenzymes B and M. F430 is restored after the catalytic cycle, whereas coenzymes B and M act as group-transferring cofactors.

F430 consists of a tetrapyrrole ring with a nickel atom (similar to the iron atom in heme) coordinated by the four nitrogen atoms and bound to a glutamine residue of the target enzyme. Compared to other tetrapyrrole molecules, it has fewer double bonds (Ermler, 2005).

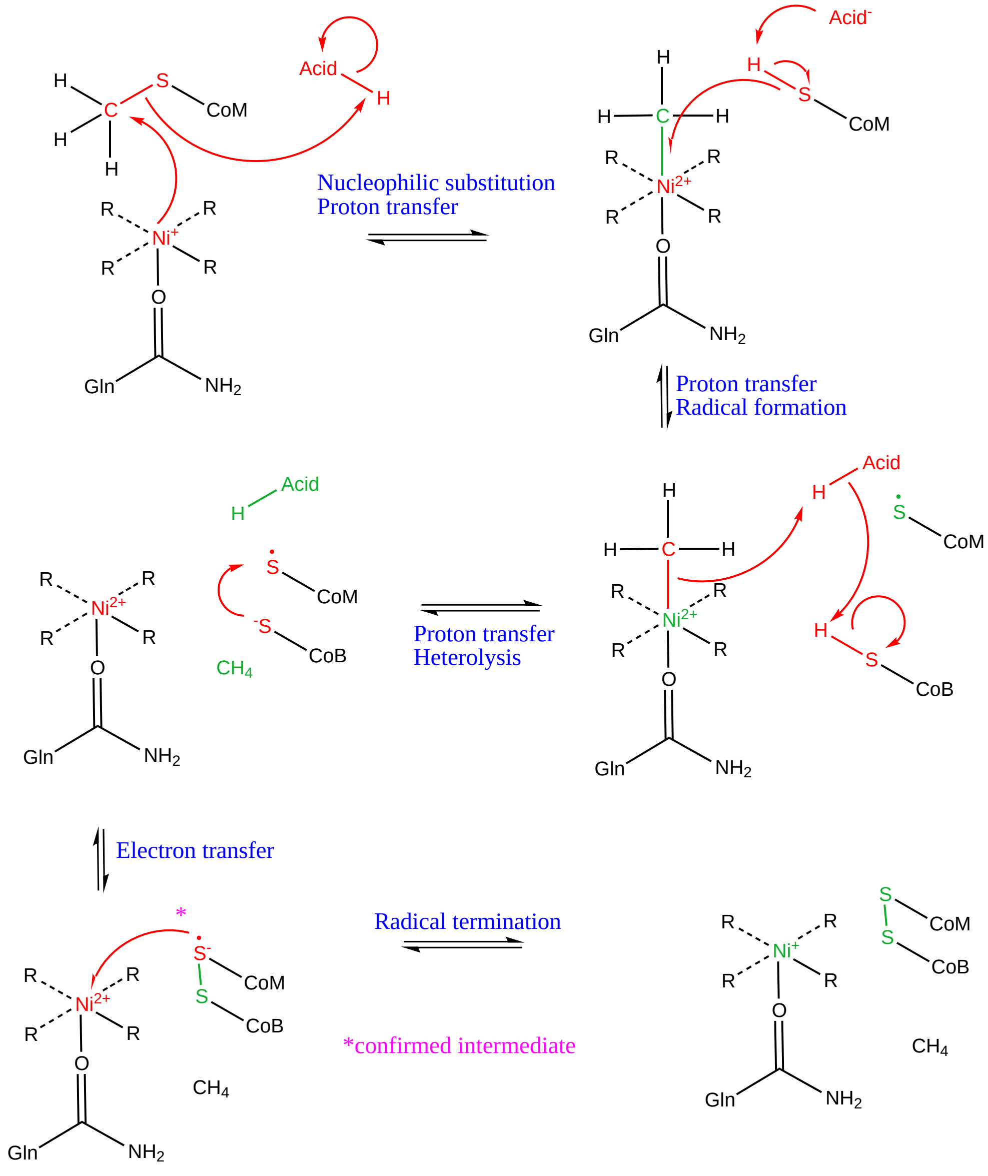
Coenzyme M (CoM) acts as the methyl-group donor in the methanogenesis reaction (Buckel & Golding, 2006b). CoM enters the methanogenesis reaction with a methyl group bound to the terminal sulfur atom (see figure 3.16). After the transfer of the methyl group to F430, coenzyme M is a radical with an unpaired electron at the terminal sulfur atom, that later reacts with the negatively charged coenzyme B sulfur to form a disulfane radical, which transfers its unpaired electron back to the nickel atom of F430 where it originated.

Neither CoM, nor coenzyme B (CoB) occur in human metabolism (Ermler, 2005). CoB acts as a base in the methanogenesis reaction (see figure 3.16). It binds the CoM radical to form a disulfane radical. The CoB-S-S-CoM is then recycled by another enzyme (Buckel & Golding, 2006b).

The CoFactor database only contains the coenzyme-B sulfoethylthiotransferase reaction for each of these three cofactors. It occurs only in archaea.

3.2.20 Heme

Heme is a generic name comprising several prosthetic groups. They all have an iron atom, coordinated by four nitrogen atoms from a porphyrin ring and, in human cells, they are generally biosynthesised. In heme oxygenase (1 and 2, E.C. 1.14.99.3), heme has a histidine as its fifth ligand and water molecule as distal ligand (Zhu & Silverman, 2008). A comprehensive review of all heme molecules



CH_3-CoM + CoB-H + F43 <==> CoM-CoB + CH_4 + F43

Figure 3.16: The mechanism of coenzyme-B sulfoethylthiotransferase (E.C. 2.8.4.1. see also MACiE M0156).

has been published by (Smith *et al.*, 2010). The authors have kindly agreed to allow me to print a figure from their publication, which is shown in figure 3.17.

ChEBI defines that "heme is any tetrapyrrolic chelate of iron". The cofactor is essential for all aerobic organisms (Furuyama *et al.*, 2007). Heme can act as a cofactor in enzymes but also as a prosthetic group in non-enzyme proteins. The cofactor functions comprise:

- serving as a source of electrons
- electron transport in the respiratory chain and photosynthesis (cytochromes, Tripathy *et al.*, 2010)
- mediating the oxidative metabolism and detoxification of xenobiotics and drugs in cytochrome P450s (Zhu & Silverman, 2008)
- catalysing the monooxygenation of substrates: R-H → R-OH (Zhu & Silverman, 2008)
- detoxification of hydrogen peroxide (peroxidase, catalase, Mendel *et al.*, 2007)

Heme's non-cofactor functions in non-enzyme proteins comprise:

- transferring and storage of gases, e.g. in hemoglobins (Mendel *et al.*, 2007)
- regulatory roles:
  - oxygen stress sensor: heme induces heme oxygenase 1, which produces biliverdin and later bilirubin (antioxidants and heme degradation products). Heme therefore functions as an oxidative stress sensor. It also suppresses the first step of heme synthesis (Unno *et al.*, 2007)
  - regulating the formation of the cell signalling molecule NO (Zhu & Silverman, 2008)
  - regulating K^+ channels (Furuyama *et al.*, 2007)

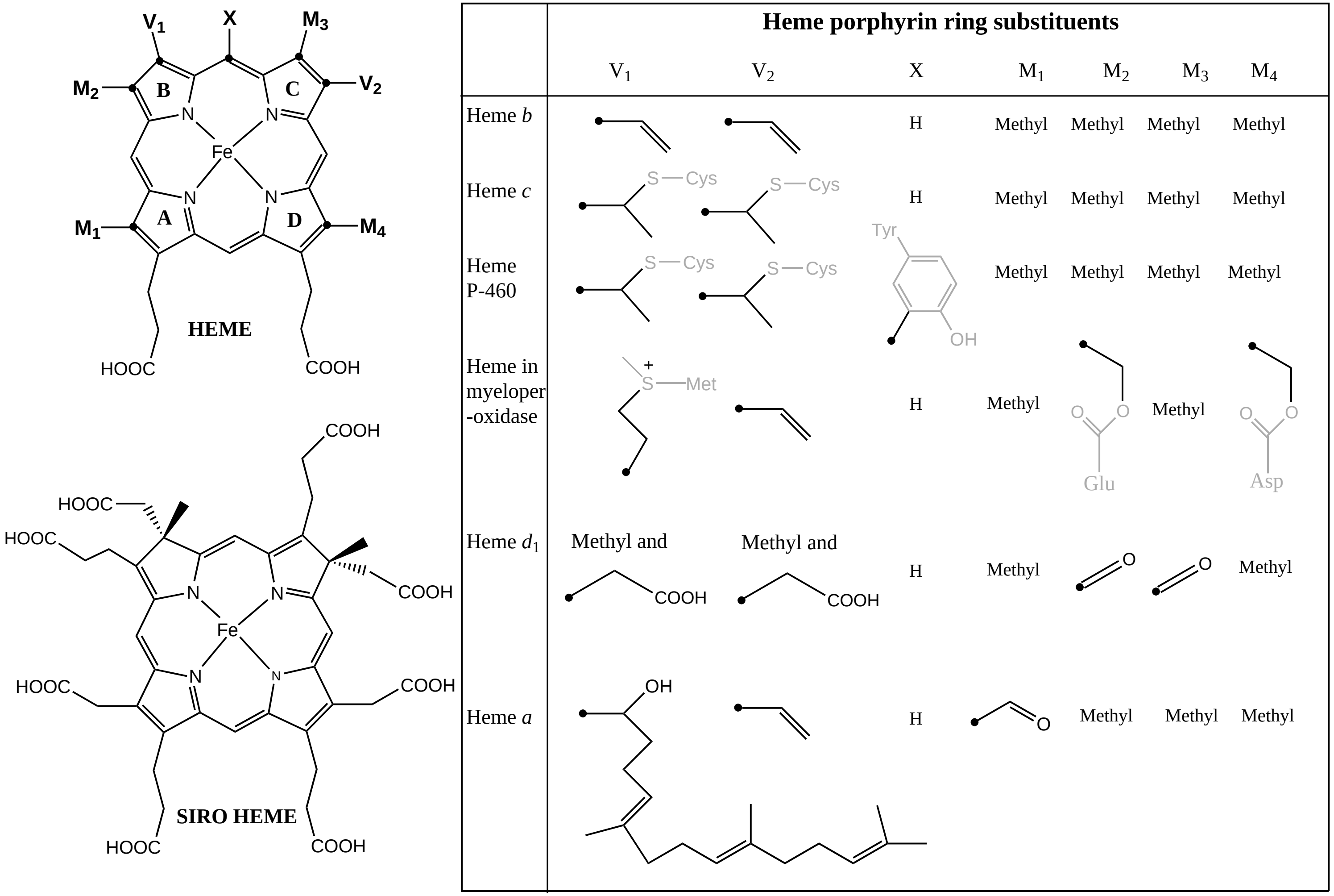


Figure 3.17: Structures of the different heme group variants. The porphyrin ring skeleton is shown at the top left and the various substituents are given in the table on the right. Atoms in the substituents bonding to the porphyrin ring (black spheres), protein atoms (gray). Figure unchanged from Smith *et al.* (2010).

Often, heme is bound to a protein, which acts as an electron carrier substrate in the reaction. Mostly, the iron alone absorbs the electron changes. In the enzyme chloride peroxidase (E.C. 1.11.1.10, see MACiE mechanism M250) however, the porphyrin ring system formally lacks one electron and therefore exists in a radical state (Woggon *et al.*, 2001).

Heme is a lipid soluble molecule. In the majority of hemoproteins, heme is not covalently attached to the apoprotein (Mendel *et al.*, 2007) and recognised by the sequence motif CP (Unno *et al.*, 2007). However, in some cases the cofactor is covalently linked to the protein (see figure 3.17).

In mammals, eight enzymes are involved in heme biosynthesis (Furuyama *et al.*, 2007). The first enzyme in mammal heme biosynthesis (5-aminolevulinate synthase) uses PLP as a cofactor (E.C. 2.3.1.7). Hydroxymethylbilane synthase (E.C. 2.5.1.61) uses dipyrromethane, a cofactor that exclusively serves as a seed for the meta-ring in heme molecules (see section 3.2.21).

Acute intermittent porphyria is an autosomal dominant mutation of hydroxymethylbilane synthase. Precipitating factors like drugs, hormones, alcohol, starvation, stresses induce ALAS-1 (non-specific 5-aminolevulinate synthase) expression in the liver, which leads to accumulation of 5-aminolevulinic acid and prophobilinogen. That, in turn, leads to symptoms like abdominal pain, vomiting, nausea. Therefore, ALAS-1 is considered a drug target. A mutation in ALAS-1 can be treated by supplying heme to repress ALAS-1. The mutation of ALAS-2 (an isozyme of ALAS-1) is X-linked and causes X-linked sideroblastic anemia, which is treated by supplying pyridoxal because ALAS-2 requires PLP (Furuyama *et al.*, 2007).

Siroheme (SRM) acts as a cofactor in the reduction of sulfate and nitrate (Tripathy *et al.*, 2010).

Heme-dependant enzymes occur in all three domains of life. The vast majority (94%, 122 E.C. numbers) of the heme-dependant enzymes in the CoFactor database are oxidoreductases.

3.2.21 Dipyrromethane

Dipyrromethane (DPM) is a prosthetic group that is biosynthesised in human cells. The cofactor acts as a nucleophile (Shoolingin-Jordan, 1995) and as a seed for heme synthesis (Warren & Scott, 1990). The enzyme hydroxymethylbilane synthase (E.C. 2.5.1.61) is the only enzyme that is known to require the dipyrromethane cofactor (Shoolingin-Jordan, 1995).

DPM is composed of two molecules of the trisubstituted pyrrole porphobilinogen (Warren & Scott, 1990) linked by a CH bridge (Shoolingin-Jordan, 1995). It is attached covalently and irreversibly to the enzyme *via* a cysteine's sulfur atom. The free α position is used to bind the first porphobilinogen substrate ring (Shoolingin-Jordan, 1995). The heme biosynthetic pathway contains the enzyme that uses and biosynthesises the dipyrromethane cofactor (Warren & Scott, 1990).

DPM acts as a reaction primer for the assembly of the tetrapyrrole macrocycle for heme biosynthesis (Shoolingin-Jordan, 1995). Since it is composed of two pyrrole units and the product of the reaction (tetrapyrrole) is composed of four pyrroles, the reaction of porphobilinogen deaminase proceeds by a stepwise addition of pyrroles to the two seeding ones from DPM. When six pyrroles are linked, water cleaves the tetrapyrrole off and restores the cofactor. Labelling experiments have shown that DPM is not exchanged (Shoolingin-Jordan, 1995). The mechanism is shown in MACiE entry M0206.

Specific biosynthetic enzymes are not required for DPM synthesis. The porphobilinogen deaminase apo-enzyme assembles it from two molecules of porphobilinogen, which is also the substrate of the enzyme. Porphobilinogen, in turn, is biosynthesised from either succinyl-CoA or glutamate, over the intermediate molecule 5-aminolaevulinic acid (see heme biosynthetic pathway, Shoolingin-Jordan, 1995).

DPM occurs in all three domains of life, even though there is only one known DPM-dependant enzyme reaction, which is a transferase.

3.2.22 Pyrroloquinoline quinone

Pyrroloquinoline quinone (PQQ) is a prosthetic group in some enzymes. The cofactor catalyses electron transfers, like all quinone cofactors (Duine *et al.*, 1990).

All known enzymes that utilise PQQ are dehydrogenases.

PQQ has a flat tricyclic ring structure. Its core scaffold is composed of a quinoline and a pyrrole portion. The three carbolic acid groups make the compound very water-soluble in contrast to the lipid-soluble ubiquinone. The two keto groups (fully oxidised or quinone state) are the functional site of the cofactor and can also adapt a semi-quinone (half-reduced) and a quinol (reduced) state. The molecule is heat-, acid-, and photo-stable and has three possible metal liganding sites (Davidson, 2004). The two carbonyl groups of the quinone ring are involved in proton-, electron- and/or hydride-transfers.

Tyrosine and glutamic acid have been shown to be precursors of PQQ in bacteria (Duine *et al.*, 1990). PQQ was proposed to be a vitamin for humans, but the key evidence for this claim was shown to be faulty (Holscher *et al.*, 2009). Claims that mammals have PQQ-dependant enzymes have been shown to be not substantial (Felton & Anthony, 2005). The only confirmed occurrence of this cofactor is therefore in prokaryotes, lower eukaryotes and plants.

In *Gluonobacter oxidans*, there are seven putative PQQ-containing dehydrogenases, which are thought to be involved in rapid oxidation of sugars and sugar alcohols (Holscher *et al.*, 2009).

In the CoFactor database, there are eight PQQ-dependant enzymes listed, all of which are oxidoreductases. In SwissProt none of these have been annotated as enzymes encoded in archaea, but they do occur in bacteria and eukarya.

3.2.23 Topaquinone

Topaquinone (TPQ) is a polypeptide-derived prosthetic group involved in amino acid metabolism and is biosynthesised by human cells. The only known enzyme to use topaquinone are the copper amine oxidases (formerly EC 1.4.3.6, now 1.4.3.21 and 1.4.3.22). In these enzymes, TPQ catalyses the oxidation of amines. Quinones complement the mostly nucleophilic repertoire of amino acid side chains by an electrophilic function (Dubois & Klinman, 2005). TPQ is further the only known para-quinone residue. The ortho-quinones are handled separately (see section 3.2.24).

The enzyme copper amine oxidase catalyses not only the oxidative deamination of primary amines with the help of the TPQ cofactor, but also the biosynthesis of this cofactor from a conserved backbone tyrosine residue and dioxygen (Prabhakar & Siegbahn, 2004). The latter one happens only once and the enzyme returns to this state after each catalytic cycle.

The mechanism of the actual reaction is a ping-pong mechanism, and it is separated into two halves (Prabhakar & Siegbahn, 2004, similar to thiamine diphosphate mechanism, see section 3.2.1). In the first (reductive) half-reaction, the substrate reduces TPQ to produce an aminophenol species, whereas molecular oxygen reoxidises this species in the second (oxidative) half. A copper atom is present in all copper amide oxidases, which is coordinated by three conserved histidine side chains and two water molecules (Prabhakar & Siegbahn, 2004).

The half life of the monooxygenation step in TPQ formation is 9 minutes and therefore very slow compared to the cofactor reoxidation during catalytic turnover (Dubois & Klinman, 2005).

The CoFactor database contains two TPQ-dependant enzymes, both of which are oxidoreductases. Like for PQQ, none of these have been annotated as enzymes encoded in archaea in SwissProt, but the reactions do occur in bacteria and eukarya.

3.2.24 Orthoquinone residues

The orthoquinone residues (OQs) are a group of polypeptide-derived prosthetic groups. Three orthoquinones have been found as cofactors in bacterial dehydrogenases (amino acid metabolism), in which a tyrosine or tryptophan residue in the active site of an enzyme has been dioxygenated and cross-linked to another residue from the same chain, namely tryptophan (TTQ), lysine (LTQ) or cysteine (CTQ), to form an orthoquinone species. OQs are residues, which act as electron donors/acceptors in the aforementioned bacterial dehydrogenase reactions (Mure, 2004).

The quinone cofactors catalyse a two-electron two-proton reaction (dehydrogenases, Dubois & Klinman, 2005). They form part of an electron sink, much like PLP (Klinman, 2001). TTQ (see picture in table 1.1) and TPQ (see section

3.2.23) oxidise amines in amine dehydrogenases, whereas LTQ catalyses the oxidation of the ω-amino groups of specific lysine residues in collagen and elastin (Duine, 2001) to generate cross-links in these proteins. LTQ is structurally related to TPQ. It is also derived from a tyrosine residue, but contrary to LTQ, the quinone in TPQ is not cross-linked to another residue. LTQ, TTQ and CTQ may depend on this covalent link between amino acids in order to restrict the conformational mobility for catalytic turnover. TPQ biosynthesis depends critically on ring mobility (Klinman, 2001).

The mechanism of LTQ-containing lysyl oxidase is not well characterised. Mechanistic studies indicate that the reaction follows a ping-pong mechanism and that a copper atom is not catalytically important but maintains the structural integrity of LTQ or the protein (Mure, 2004). All known structures of these cofactors have an aspartic acid residue in the active site, which is thought to act as the catalytic base for proton abstraction (Duine, 2001).

The only mechanism in MACiE with orthoquinone residues is the TTQ-dependant amine dehydrogenase (E.C. 1.4.99.3, M0013). TPQ and LTQ are reoxidised by molecular oxygen, whereas TTQ and CTQ depend on an external electron acceptor (Dubois & Klinman, 2005) and must exclude molecular oxygen in order to direct the electrons to the external acceptor. The latter two orthoquinone cofactors do not co-appear with a copper centre, unlike LTQ and TPQ (Duine, 2001). Orthoquinone residues are a way to complement the mostly nucleophilic catalytic capacity of enzymes with an electrophilic function (Dubois & Klinman, 2005).

Interestingly, neither sequence nor structural homology has been detected among the enzymes using the different orthoquinone cofactors and TPQ. Nevertheless, their chemical mechanisms are expected to have common features, since they all have primary amines as substrates, with which they can form covalent adducts (Klinman, 2001).

CTQ is a cofactor of a bacterial amine dehydrogenase (E.C. 1.4.99.3, e.g. in *Paracoccus denitrifians*). This enzyme contains two additional heme C cofactors, which pass the two electrons from the substrate onto an external acceptor (Tsai *et al.*, 2009).

The orthoquinone cofactors are biosynthesised by cross-linkage of backbone amino acids. Although PQQ also has a catalytically active orthoquinone portion, it differs from TTQ, LTQ and CTQ in that it is biosynthesised outside the target enzyme (van der Palen *et al.*, 1995). Biosynthesis of LTQ (and TPQ) has been proposed to depend on the copper atom in the protein, whereas the biosynthesis of LTQ and TTQ requires further proteins (Dubois & Klinman, 2005).

In methylamine dehydrogenase (E.C. 1.4.99.3), which contains TTQ as a cofactor, the mauJ and mauG genes have been postulated to be involved in TTQ biosynthesis (Pearson *et al.*, 2004). MauG has been shown to catalyse the covalent linkage between the two tryptophan residues and the incorporation of the second oxygen atom into the ring of TTQ (Sun *et al.*, 2003). CTQ most likely also depends on external proteins for cofactor biosynthesis (Duine, 2001).

The CoFactor database lists two OQ-dependant enzymes, both of which are oxidoreductases. In SwissProt, these enzymes are only annotated in protein sequences encoded in bacterial genomes.

3.2.25 Lipoic acid

Lipoic acid (LA) is a prosthetic group, which is covalently attached to a specific lysine residue in the lipoyl domains of the pyruvate dehydrogenase multienzyme complex.

The cofactor is a sulfur-containing carboxylic acid. The two sulfur atoms are located in a five-membered ring in adjacent positions (see ChEBI entry CHEBI:43796). In enzymes it is covalently attached to the ϵ-amino group of a lysine residue, where it provides substrate binding and mobility (swinging arm, range about 14Å). The five-membered heterocycle can open between the two S-atoms (dihydrolipoyllysine, Fries *et al.*, 2003 and Reed, 1974). The best-known enzyme that depends on lipoic acid is the pyruvate dehydrogenase complex, which forms part of the centre of aerobic metabolism: the citrate cycle (see KEGG pathway ec00020).

Octanoic acid is a physiological precursor of lipoic acid. The two sulfur atoms are introduced sequentially. This introduction is mechanistically similar to the same procedure in biotin. Genetic investigations support this hypothesis. The sulfur donor is likely to be a iron sulfur cluster, like in biotin synthase. Lipoic acid is synthesised in many organisms, among them are bacteria, plants, yeast and animals (Marquet *et al.*, 2001).

The CoFactor database lists six LPA-using enzymes, five of which are oxidoreductases. These enzymes occur in all three domains of life.

3.3 Technical implementation

The CoFactor database is implemented and maintained using MySQL (initially version 5.0.41) and the web pages are generated with PHP (initially version 5.1.6). Figure 3.18 shows the database schema.

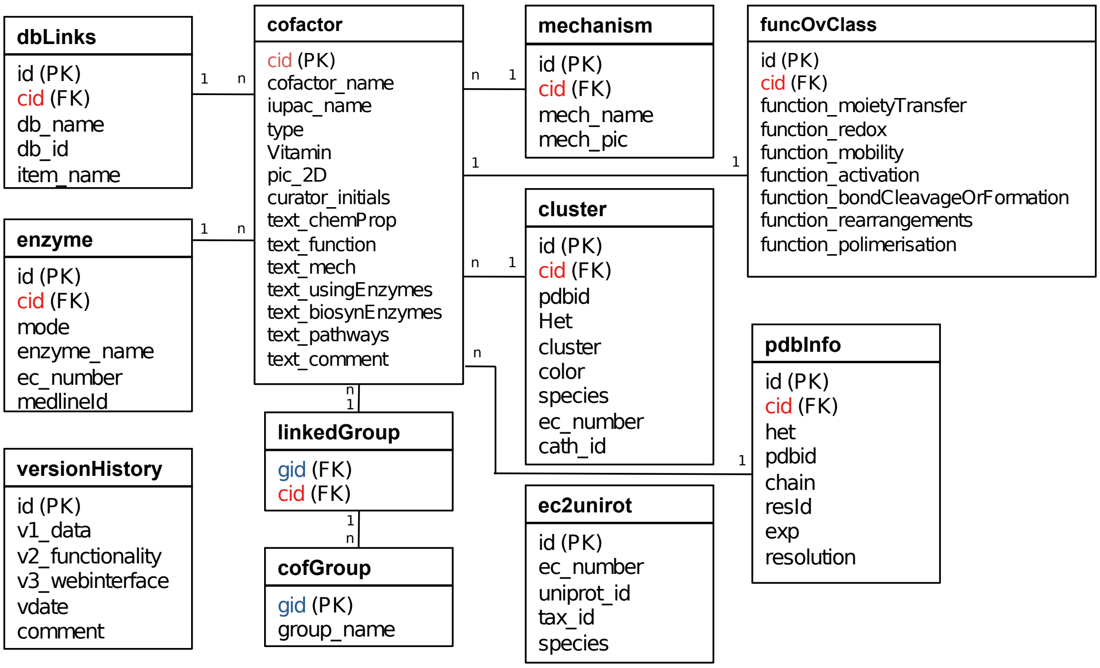


Figure 3.18: Database schema of the CoFactor database. PK: primary key; FK: foreign key.

The database schema shows all the tables in the CoFactor database and how they relate to each other. The central table is called "cofactor" and its most important column is the cofactor identifier (cid), which is used in most of the other tables to identify the cofactor that each piece of information refers to. The lines between the tables indicate that there is a reference from one table to the other one. The numbers on these lines quantify this relationship. For example, the one-to-many relationship (described by the "1" and the "n" on the line connecting the two tables) between the cofactor and the enzyme table denotes that a record in the enzyme table always refers to one specific entry in the cofactor tables, whereas an entry in the cofactor table may be referenced in many records in the enzyme table. The database comprises 11 tables. As mentioned above, the cofactor table is the central table and holds the cid. Additionally, a cofactor's common name, IUPAC name, its type (*i.e.* coenzyme or prosthetic group), it's role as a vitamin for humans and a link to a picture of its two-dimensional structure are stored here. The fields starting with "text_" hold the manually curated information. To identify possibly different curators in the future, another column ("curator_initials") is provided. PHP-based curator pages have also been created to facilitate the data entering process for curators.

The "dbLinks" table holds links to various other databases, including IntEnz (*via* the E.C. number) ChEBI, KEGG COMPOUND, PDBeChem, CATH domains, MACiE entries, PROCOGNATE (*via* the HET code) and UniProt sequences.

The "enzyme" table stores information about enzymes, such as an enzyme's name and E.C. number, as well as its mode of relationship with the cofactor it references, *i.e.* cofactor-using, cofactor-synthesising, cofactor-recycling, or – if no other information is available – cofactor-binding. The former three modes of relationship are not mutually exclusive. The field "medlineID" holds either the medline identifier of the publication that this piece of information was manually extracted from, or it references the database that was queried to obtain the information.

The tables "cofGroup" and "linkedGroup" classify all the cofactors into the groups flavin-like, phosphate-containing, nucleotide-containing, sulfur-containing, pterin-like, porphyrin-like, fat-soluble and quinone-like cofactors.

The "mechanism" table stores a link to the image depicting each of the mechanisms associated with a cofactor.

The table "cluster" stores all of the information necessary to fill the annotation on the superposition pages. For each of the structures in the applets, the table contains one row. This row stores the cid of the cofactor, the PDB code it was extracted from, the HET code, the cluster number and colour, the biological species, the E.C. number (if available) and the CATH code(s) in this PDB structure that the cofactor binds to (from PROCOGNATE if available).

The “pdbInfo” table additionally holds information that applies to the whole structure, such as the resolution, the experiment type (NMR or X-ray) and the original residue identifier of the cofactor in the structure coordinate file.

The “ec2uniprot” table stores which E.C. numbers occur in which species. Please note that this table is not explicitly linked to the cofactor table, due to the possibility of alternative (cofactor-independant) mechanisms of cofactor-dependant enzyme reactions.

The “funcOvClass” table stores the class membership of the classes defined in figure 4.14 in Chapter 4.

Finally, the “versionHistory” table will be used in the future to store the changes made to the database from each version to the next. The first release version is 1.0.0. Adding new data will increase the first of these numbers (“v1_data”), adding functionality will increase the second one (“v2_functionality”), and improving the web interface will increase the third number (“v3_webinterface”). The date of the update and a comment describing the changes will be stored too.

The size of the MySQL data base is 6.5 MB and the size of the data (including images, Jmol scripts and PHP scripts) is 26.9 MB in CoFactor version 1.0.0.

3.4 Discussion and relevance

The goal of the CoFactor database is to provide the user with an overview of all organic enzyme cofactors and to integrate that overview with information about the enzymes that use them.

Overall, the database contains 27 organic enzyme cofactors: nine coenzymes, eleven prosthetic groups, four that act as both and three polypeptide-derived cofactors.

Browsing the CoFactor database web pages, it becomes clear that the cofactors are used to highly varying degrees and that the E.C. profile strongly depends on the type of cofactor. While some only occur in one or a handful of reactions (e.g. factor F430 or dipyrromethane), others assist in the catalysis of hundreds of enzyme reactions (e.g. NAD or FAD). Similarly, while some cofactors only occur in a very small range of biological species (e.g. coenzyme M and factor F430), others are essential for all forms of life (e.g. thiamine diphosphate). Twelve of the cofactors are (biosynthesised from) vitamins for humans, *i.e.* their availability depends on nutrition and they cannot be fully biosynthesised. Therefore, many cofactors are linked to diseases. However, this does not only include deficiency diseases caused by the lack of, or reduced uptake of the vitamin (e.g. scurvy for vitamin C deficiency), but also genetically caused diseases where mutation in the biosynthesis or uptake systems cause symptoms (e.g. biotin or glutathione).

The diversity of these molecules raises the question whether there are intrinsic groupings among them and if so, whether these are defined by the molecules' physicochemical properties, their two-dimensional structural scaffolds, or their specific three-dimensional conformations. It further raises the question if they can be functionally classified, to what extent cofactors are generally used in the different enzyme classes and whether they contribute new catalytic capabilities to the enzymatic toolkit. All of these questions will be investigated in chapters 4 and 5.

It is worth noting that the CoFactor database is expected to be comparatively easy to maintain. As most of the cofactor molecules have been discovered several decades ago, no exponential growth of entries is expected, although the occasional new discovery is not impossible. However, the expected growth rate is far below the one observed for the large nucleotide or amino acid sequence databases. The manually curated part should be updated in regular intervals, ideally once a year, and in case the project is continued, the functionality may be extended. The data integration part is largely automated and can thus be updated automatically. All of the data in the CoFactor database provides the user with a glimpse of this fascinating group of molecules and an entry point for further research.

Chapter 4

The structures and physicochemical properties of organic cofactors in biocatalysis

After having established a well-defined data set and gained an overview of each organic cofactor (chapter 3), this chapter covers the data analyses that have been performed on the one- and two-dimensional cofactor data as well as the classification and clustering results. This information is necessary to broaden the understanding of the organic cofactors as a group of molecules and as catalytic entities.

In the first section, the data analysis based on one-dimensional descriptors of the cofactor molecules will be presented and in the second section the results from the clustering of the molecules' two-dimensional chemical structures will be covered. Finally, the molecular function of each of the cofactors will be discussed at an overall enzyme reaction level as well as on a more detailed stepwise mechanism level. Some of the analyses in this chapter have been published (Fischer *et al.*, 2010a).

4.1 One-dimensional analysis of molecular descriptors

In order to better understand the physicochemical properties of the organic cofactors, eight common descriptor variables have been calculated for each of the molecules in the cofactor and metabolite data sets described in section 2.2. The molecular property descriptor variables were computed for all the cofactors and metabolites in the respective data sets. Interesting variables include those used in Lipinski's rule of five (Lipinski *et al.*, 2001) and a description of all eight chosen ones can be found in section 2.2.2.

These data can now be used to firstly get an overview of the physicochemical properties that the organic cofactors exhibit and secondly compare the cofactors' properties to the ones of other metabolites in living cells in order to assess whether there are significant differences between the two data sets, that might explain the catalytic properties exhibited by the cofactors, but not by the metabolites. The resulting data matrix of the cofactor descriptors is shown in table 4.1.

4.1.1 Distribution of molecular descriptors in the cofactor data set

To tackle the first objective, namely getting an overview of the physicochemical properties, a histogram is plotted for each of the eight descriptor variables in figure 4.1. Each bar is labelled with the cofactors (HET codes) that are binned in this bar.

This figure allows for visual identification of outliers and rough distribution shapes. The y-axes are not normalised so that the original scales of the variables can be seen.

Notably, B12 appears mostly at the ends of the distributions. All the descriptors with a size component (molecular weight, number of rings, number of hydrogen bond acceptors and donors and the polar surface area) show B12 in one of the two rightmost bins. Table 4.1 confirms that B12 is the largest (heaviest) cofactor in the data set, with a molecular weight of considerably over $1kDa$. As the PSA, HBA and HBD variables also measure polarity, a comparison to the

HET	mw	PSA	HBA	HBD	logP	percRotBd	numRings	atomComp
ASC	176.03	107.22	6	4	-1.41	0.5	1	0.5
B12	1328.56	470.16	27	10	5.98	0.47	11	0.31
BTN	244.09	103.73	5	3	1.45	0.35	2	0.38
COA	767.11	414.79	23	9	-0.3	0.6	3	0.56
COM	141.98	101.55	3	1	0.88	0.67	0	0.71
DPM	420.15	173.92	10	4	0.82	0.55	2	0.33
F43	903.27	299.53	19	7	2.21	0.32	10	0.31
FAD	785.16	382.55	24	9	-1.79	0.4	6	0.49
FMN	456.11	217.9	13	6	-1.56	0.42	3	0.45
GSH	307.08	197.62	9	5	-0.72	0.79	0	0.5
H4B	241.12	136.29	8	6	-0.84	0.33	2	0.47
HEA	852.35	131.62	10	3	6.42	0.42	8	0.17
HEC	618.19	94.32	8	2	0.64	0.28	8	0.19
LPA	206.04	87.9	2	1	2.79	0.5	1	0.33
MGD	740.06	443.85	23	10	-0.62	0.33	6	0.57
MQ7	648.49	34.14	2	0	14.1	0.59	2	0.04
MTE	395.01	259.23	11	6	0.32	0.31	3	0.58
NAD	664.12	338.93	21	8	-0.94	0.4	5	0.52
NAP	744.08	395.27	24	9	-0.82	0.42	5	0.56
PGD	738.04	440.22	23	8	-3.05	0.33	6	0.57
PLP	247.02	126.76	6	3	0.52	0.5	1	0.5
PQQ	330.01	174.72	10	4	0.55	0.23	3	0.42
PTE	1028.94	523.42	24	12	1.53	0.37	6	0.6
SAM	399.14	207.93	11	5	0.01	0.45	3	0.44
SRM	914.2	314.74	20	8	1.68	0.43	8	0.32
THF	473.17	223.33	14	7	-0.1	0.42	3	0.41
TP7	345.1	184.93	8	5	1.12	0.9	0	0.48
TPP	425.05	217.88	11	4	2.22	0.52	2	0.54
TPQ	195.05	97.46	5	2	0.12	0.36	1	0.36
U10	862.68	52.6	4	0	17.85	0.71	1	0.06

Table 4.1: Objects × variables data table for the cofactor physicochemical properties (see section 2.2.2). The abbreviations of the cofactors (HET codes) are defined in table 1.1. Molecular weight (mw), polar surface area (PSA) in Å^2, number of hydrogen bond acceptors (HBA) and donors (HBD), logP value (logP), percentage of rotatable bonds (percRotBd), number of rings (numRing) and the atomic composition (atomComp).

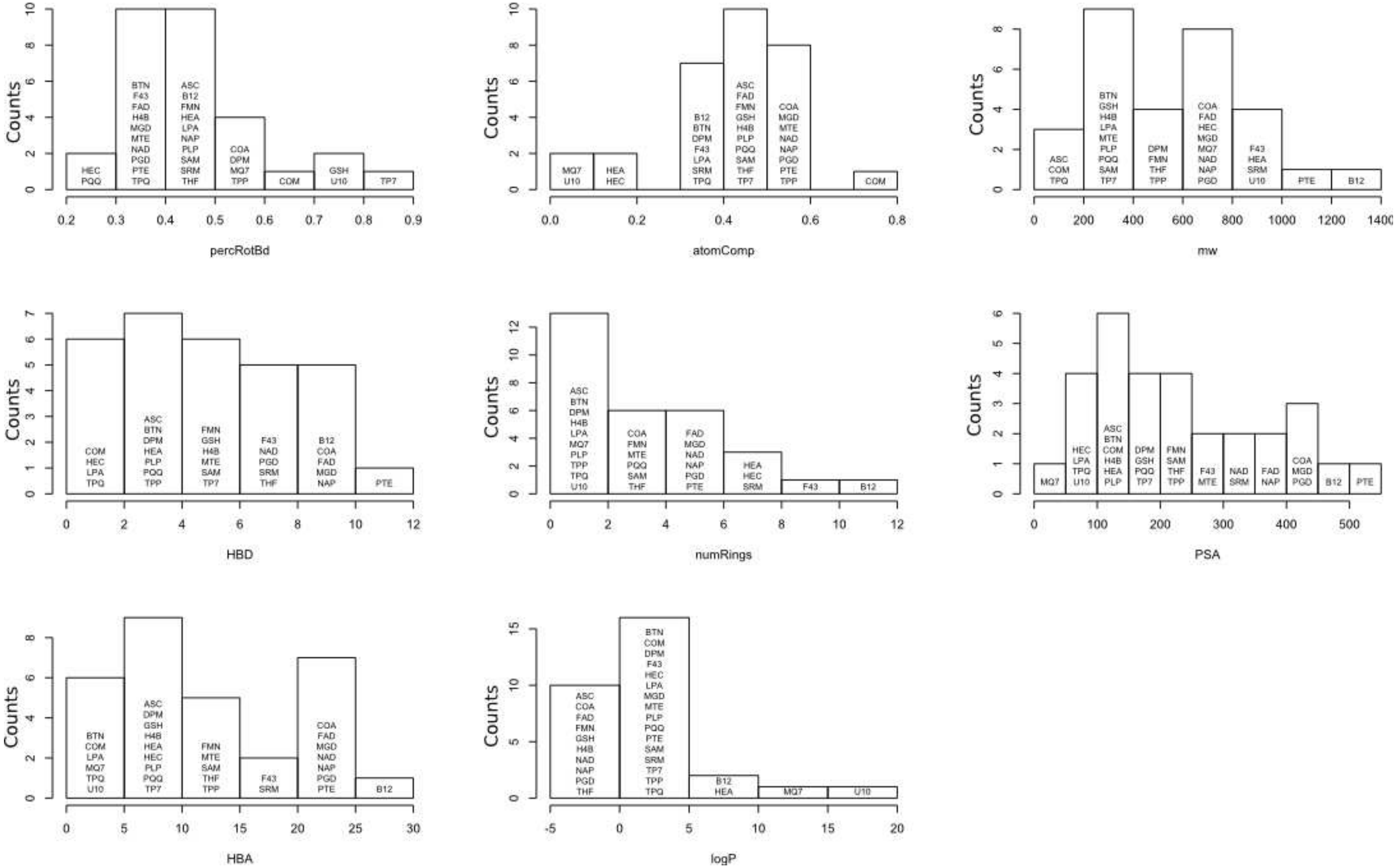


Figure 4.1: Histograms of the eight descriptor variables as defined in section 2.2.2. The y-axes give absolute counts.

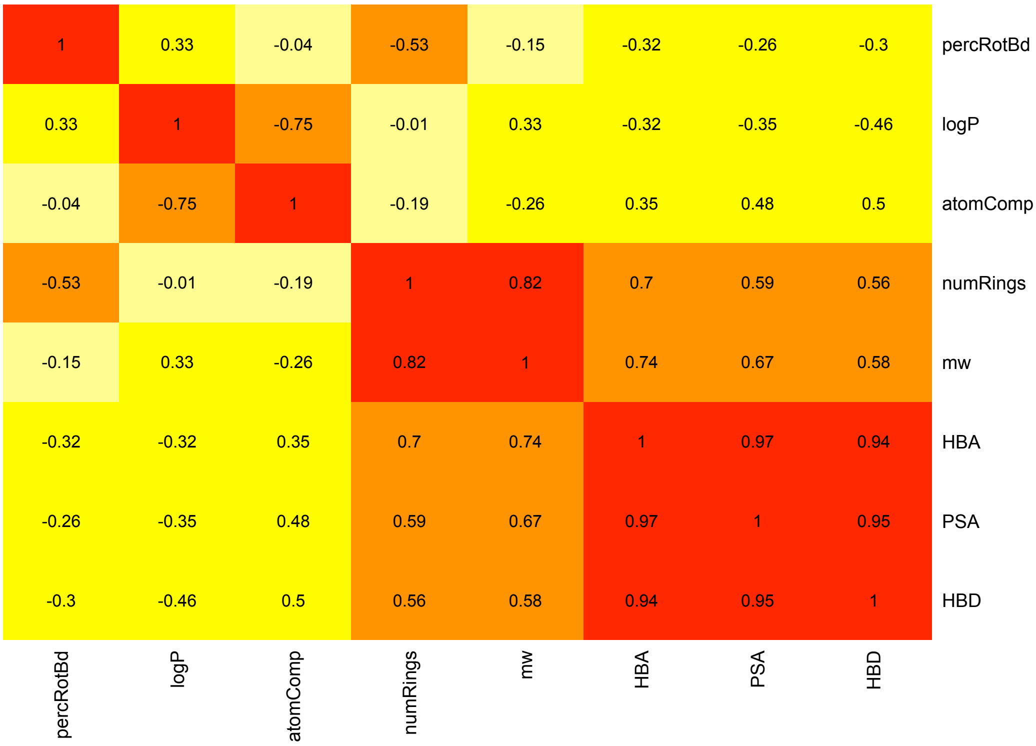
pure and size-independant polarity and hydrophobicity variables is appropriate. Interestingly, the atomic composition value for B12 (0.31) is close to the average of the data set (min=0.04, mean=0.42, max=0.71, see also figure 4.4B). B12's logP value (5.98) exhibits a similar scenario (data set: min=-3.05, mean=1.67, max=17.85), although B12 is still an outlier in this distribution (see figure 4.4B). These results show that the placement in the far right-hand side bins in the PSA, HBA and HBD histograms is mainly driven by B12's unusual size rather than its polarity.

The cofactors MQ7 and U10 also have a preference to appear at the extremes of the distributions, predominantly in the variables with a polarity/hydrophobicity component. They do not appear in the HBD histogram since they have no hydrogen bond acceptors at all (cf. table 4.1). With 2 and 4 hydrogen bond donors, 2 and 1 rings, 34.14 and 52.60 Å^2 of polar surface area and an atomic composition of 0.04 and 0.06 respectively, MQ7 and U10 are located at the lower end of the variable ranges, whereas their logP values top the range in the data set with 14.10 and 17.85 respectively. Together, these values arise from the very hydrophobic nature of MQ7 and U10. In the context of their biological function this makes sense as MQ7 and U10 act as electron carriers in biological membranes (Berkner, 2008; Cluis *et al.*, 2007).

Having identified the unusual molecules in the data set, the second step is to evaluate the shape of the descriptor distributions visually. Figure 4.1 shows that none of the descriptor histograms adopt a perfect normal distribution. Given the small sample size, however, the distributions appear to be reasonably continuous.

Figure 4.2 shows a heat map of the pairwise correlation matrix of all eight descriptor variables. The colour coding ranges from pale yellow (low correlation) to red (high correlation).

Extreme correlations between the HBA, HBD and PSA variables are immediately obvious (0.94, 0.95 and 0.97). The molecular weight variable is highly correlated with the number of rings variable at 0.82. All remaining correlations are 0.75 or less. The correlation between the two above mentioned groups (PSA, HBA, HBD compared to mw, #rings) of descriptors are in the range of 0.56 to 0.74 and thus still considerably high. All these variables have a size component: the molecular weight is a direct size measure while the number of rings increases



	percRotBd	logP	atomComp	numRings	mw	HBA	PSA	HBD
percRotBd	1	0.33	-0.04	-0.53	-0.15	-0.32	-0.26	-0.3
logP	0.33	1	-0.75	-0.01	0.33	-0.32	-0.35	-0.46
atomComp	-0.04	-0.75	1	-0.19	-0.26	0.35	0.48	0.5
numRings	-0.53	-0.01	-0.19	1	0.82	0.7	0.59	0.56
mw	-0.15	0.33	-0.26	0.82	1	0.74	0.67	0.58
HBA	-0.32	-0.32	0.35	0.7	0.74	1	0.97	0.94
PSA	-0.26	-0.35	0.48	0.59	0.67	0.97	1	0.95
HBD	-0.3	-0.46	0.5	0.56	0.58	0.94	0.95	1

Figure 4.2: Heatmap of the correlation matrix of the cofactor descriptor matrix (see table 4.1), coloured by absolute correlation values: pale yellow [0 : 0.25], yellow]0.25 : 0.5], orange]0.5 : 0.75] and red]0.75 : 1].

with increasing size. It is worth noting that nearly all cofactors do have at least one ring system (all except COM, TP7 and GSH) and that the larger cofactors have more rings (e.g. B12 has 11). The other three descriptors (HBA, HBD and PSA) also have a polarity component.

The next highest absolute correlation of -0.75 in the heat map is between the atomic composition and the logP descriptors. While logP is a measure of the hydrophobicity of a molecule, the atomic composition is the percentage of polar heavy atoms over all heavy atoms and thus a measure for (size-independant) polarity.

The last remaining absolute correlation over 0.5 is another anticorrelation between the number of rings and the percentage rotatable bonds variables (-0.53). Thus, these two descriptors are considerably anticorrelated. This relationship is not unexpected as a high number of rings reduces the number of freely rotatable bonds in a molecule through higher interconnectivity of its atoms.

4.1.2 Distribution of molecular descriptors: comparing cofactors to other metabolites

To address the second question, namely comparing the cofactors' properties to the ones of other metabolites, each descriptor variable is plotted for both of the data sets as two superimposed distributions: the metabolites distribution is shown as a smoothed line, while the cofactor distribution is plotted as a histogram. These representations have been chosen to account for the significant difference in data set size: the cofactor data sets consists of 30 molecules whereas the metabolite data set comprises 12,548 molecules. The results are shown in figure 4.3.

In figure 4.3, the observed density distributions of the physicochemical properties of all cofactors (black) are compared to all metabolites (red). A *t*-test (see section 2.2.3) was applied to determine if there are significant differences in the means of the two data sets for each descriptor variable. Using a significance *p*-value cut-off of 0.01, these distributions highlight that cofactors are, on average, significantly more polar than the metabolites, *i.e.* have more hydrogen bond acceptors (HBAs, figure 4.3E), more hydrogen bond donors (HBDs, figure 4.3D), a higher atomic composition score (figure 4.3B), and more polar surface area (PSA, figure 4.3F). Visual inspection of figure 4.3 shows that cofactors have, on average, slightly higher molecular weights (figure 4.3A), a higher percentage of rotatable bonds (figure 4.3H) and more rings than the molecules in the metabolite data set (figure 4.3C); however, these differences are not statistically significant. The logP (figure 4.3G) values, are nearly identical for both data sets with a *p*-value of 0.558.

4.1.3 Principal component analysis of molecular cofactor descriptors

The results from the previous section start to paint a picture of the relationships between the physicochemical properties of the cofactors and their comparison to

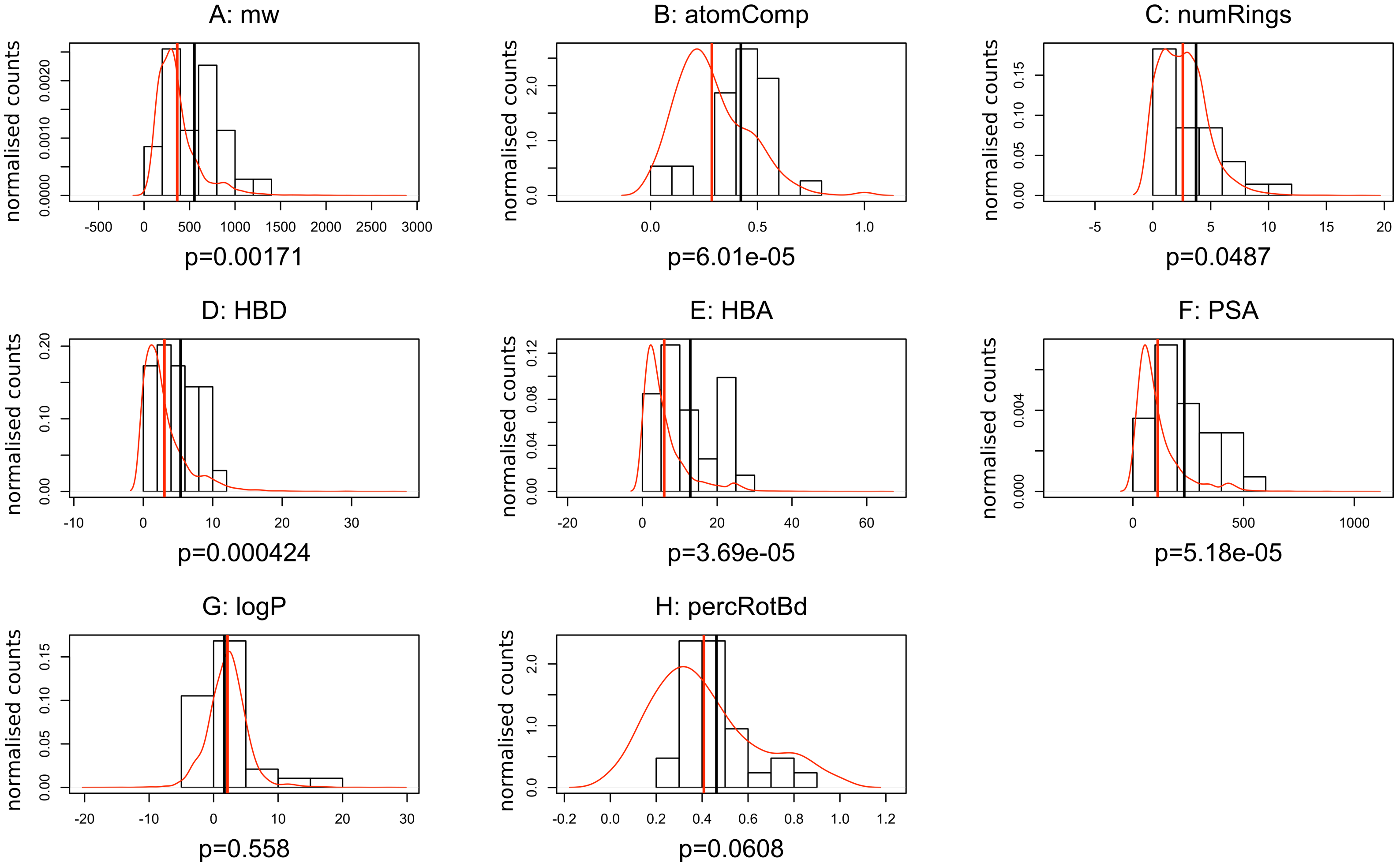


Figure 4.3: Density distribution of descriptors on 30 cofactors in our data set (black histogram) and on the metabolites data set (red curve). Averages are indicated by vertical lines (30 cofactors black and 12,548 metabolites red). Averages are considered to be significantly different if the p-value is below 0.01. Axes: units of raw data, such that the area under the graphs equals 1. (A) molecular weight, (B) atomic composition, (C) number of rings, (D) number of HBDs, (E) number of HBAs, (F) PSA, (G) logP value, (H) percentage of rotatable bonds.

those of the metabolites. To obtain a more integrated and informative view of the descriptor data, the next obvious step is to move from several univariate analyses to one multivariate view. Therefore, a principal component analysis (PCA) has been performed on the physicochemical property descriptors in table 4.1.

4.1.3.1 Preconditions of the data set

As described in the Methods chapter (section 2.2.5.3), some preconditions of the input data set need to be checked before a PCA can be performed and interpreted. Since the data are measured in different units and scales, it is necessary to normalise the data (as described in section 2.2.5.1).

Scaling and outliers

Figure 4.4 uses the boxplot method (see section 2.2.5.2) to visualise how the unscaled cofactor data set (A) and the scaled one (B) are distributed and where the outliers are.

If the unscaled data (figure 4.4A) were used to perform the PCA, the influence of variables measured in scales with a larger range (e.g. molecular weight, range [141.98 : 1328.56]) would mask the influence of the ones measured in smaller scales (e.g. HBA, range [2 : 24]). Thus, the scaled data (figure 4.4B) were used for this analysis.

Only two descriptors show outliers: the percentage of rotatable bonds (two outliers: TP7 with 0.90 and GSH with 0.79) and the logP value (four outliers: U10 with 17.85, MQ7 with 14.10, HEA with 6.42 and B12 with 5.98). As these values are correct and do not arise from an error in measurement, exclusion of these cofactors from the analysis is not justified.

Normal distribution of data

The histograms in figure 4.1 and the boxplots in figure 4.4B both visualise the distribution of the descriptor variables. As mentioned above, the variables are not necessarily normally distributed. Due to the small number of elements in the data set, however, the distributions are mostly free of outliers. Hence the analysis will be continued. As noted in the Methods chapter 2.2.5.3, it is worth reiterating that there is no guarantee that the principal components (PCs) are

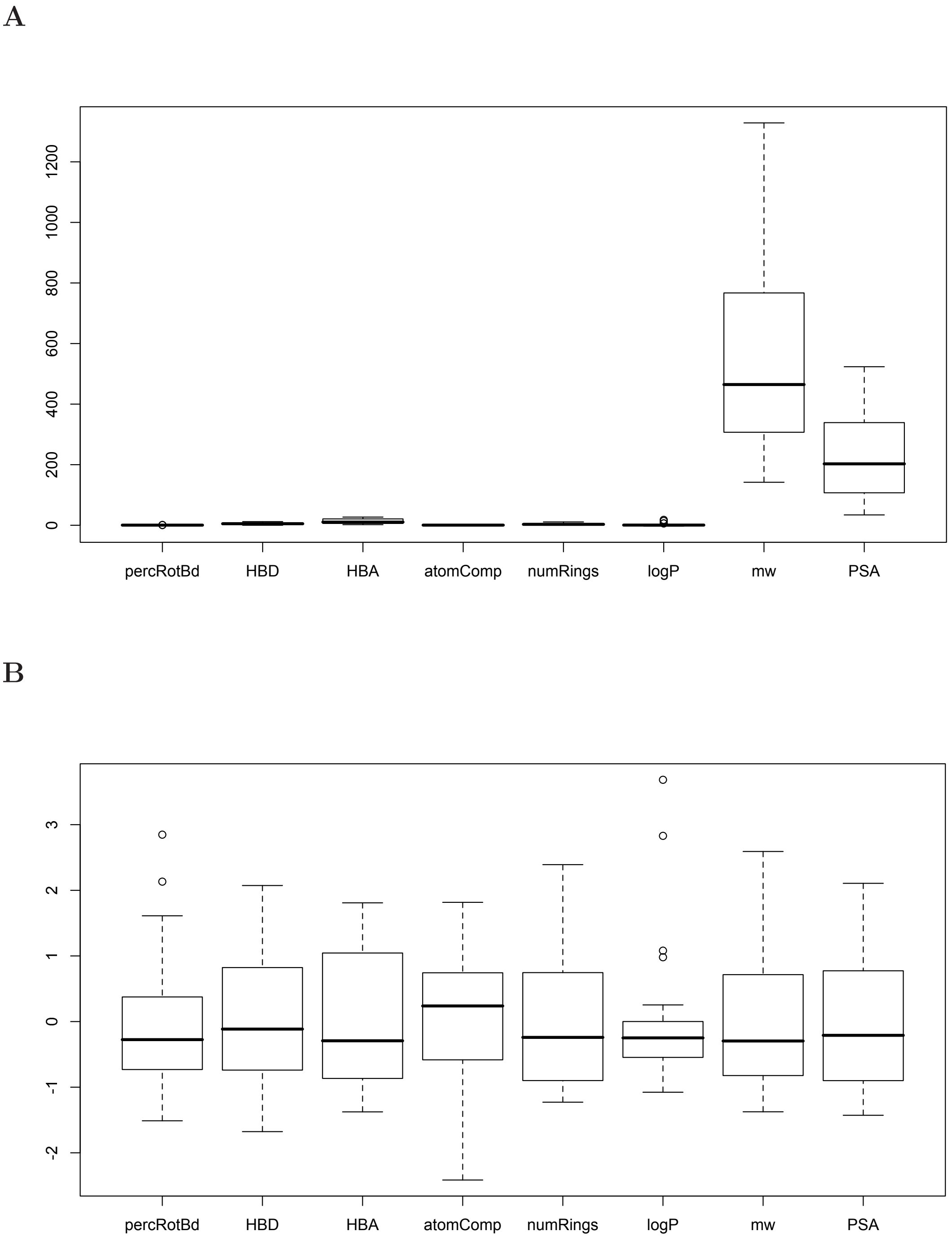


Figure 4.4: Boxplots of the eight descriptor variables in the cofactor data set (cf. table 4.1: (A) unscaled and (B) scaled data set.

good discriminatory features if the variables are not normally distributed. Yet, the PCA will still decorrelate the data.

4.1.3.2 Evaluation of PCA results

The PCA plot in figure 4.5 provides a descriptor-based representation of cofactor similarities and differences. It shows which of the descriptor variables contribute most to the variability in the data set and how these variables relate to one another. Additionally, the PCA plot allows one to visually understand and cluster the relationships between the descriptor variables and the cofactors. Figure 4.6 shows the contribution of each descriptor variable to each of the PCs and will be evaluated in detail below.

Figure 4.5 shows a “biplot” (see section 2.2.5.4) of the first two PCs, thus plotting both the cofactors as labels and the descriptor variables as vectors (arrows). Together, the first two PCs account for 81.49% of the variability in the data. Small angles (close to 0°) between the vectors indicate high correlation between variables, whilst orthogonality indicates independence, and large angles (close to 180°) indicate anticorrelation of the variables. The length of the vectors quantifies the amount of contribution of each variable in these first two PCs.

It is immediately obvious that cofactors vary most in polarity and size: the first PC is dominated by the PSA, number of HBA, and number of HBD variables (eigenvalues $\geq$0.8, “polarity”) as well as the molecular weight and the number of rings (eigenvalues $\geq$0.7, “size”), and the second PC by the atomic composition and the hydrophobicity (logP) variable (eigenvalues $\geq$0.8, “polarity”). It is worth noting that there are anticorrelations between the percentage number of rotatable bonds (percRotBd) and the HBA variables, as well as between the logP and the atomic composition (atomComp) variables. The flexibility of the molecules is a further contributing factor to the variability of cofactors, but only ranks third, with the percentage of rotatable bonds being the highest contributor to PC three (eigenvalues $\geq$0.8, “flexibility”). The first three PCs cumulatively account for 94.5% of the variability in the data.

It is worth taking a closer look at the loadings of the PCA in order to understand, which variables contribute to what extent to each principal component.

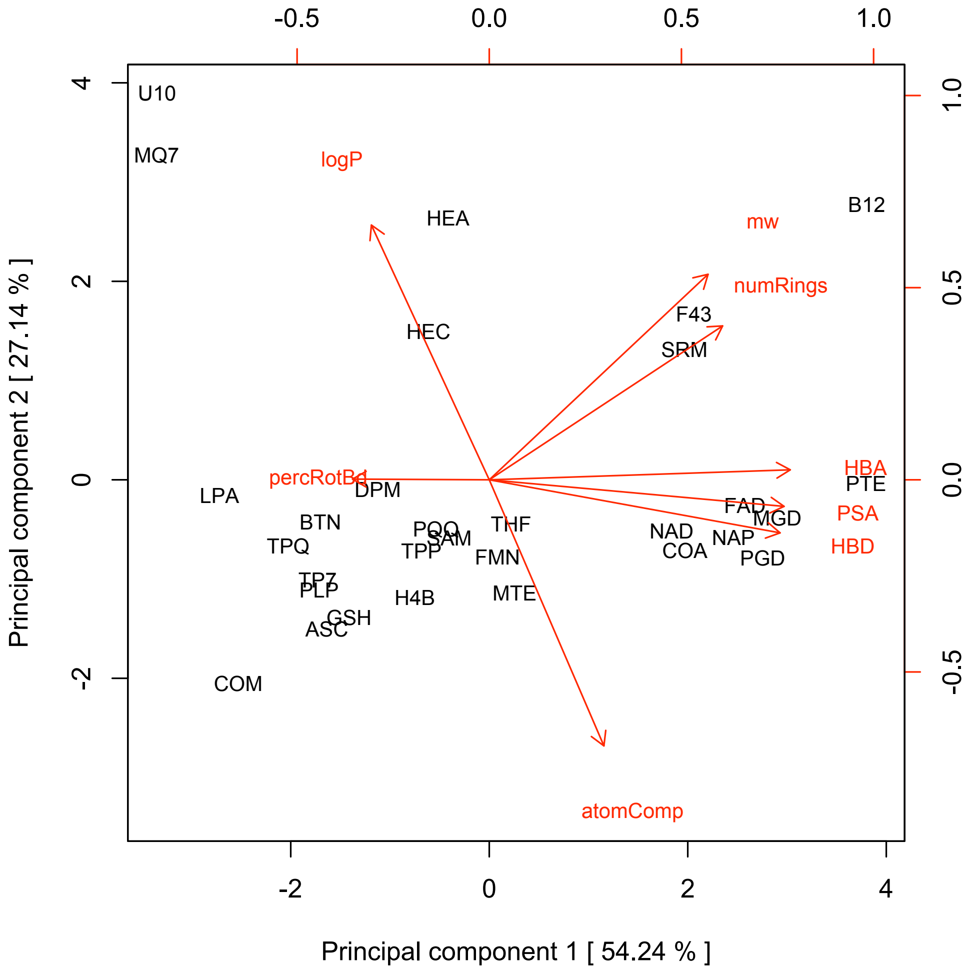


Figure 4.5: Biplot of the PCA. The red arrows show the projection of the input variables onto the space, which is spanned by the first two principal components. The length of each red vector represents the amount of contribution to this space. The black labels show the position of the cofactors within this space. All PCA plots have been generated using the FactoMineR13 and ade414 packages of the R-project. The input data are scaled.

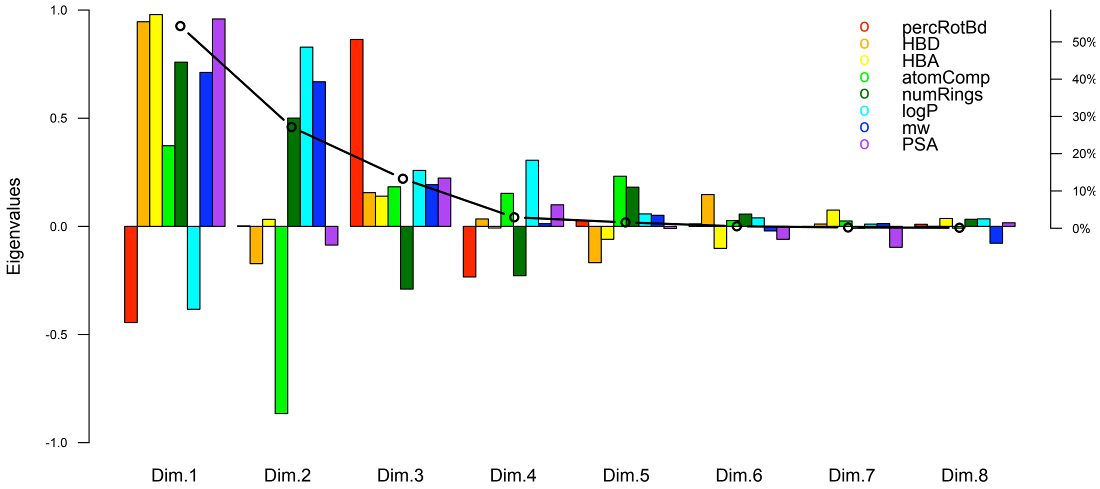


Figure 4.6: Loadings of the PCA. The eigenvalues (coloured bars, left axis) are scaled by the cumulative percentage contribution (black circles and line, right scale) to the total variability in the data.

Figure 4.6 shows the contribution of each of the eight descriptors to each of their eight PCs. The height of each bar has been scaled by its PC's contribution to the overall variability in the data. Since the relationship between the eigenvalues and the variability is squared, this implies that the sum of the squares of all the bars of each descriptor variables equals 1. This scaling visualises the importance of each descriptor variable in relation to the importance of the PC. The sign of the eigenvalue indicates correlation (+) or anticorrelation (-) to the PC. The height of the bar indicates the extent of the correlation.

Figure 4.6 immediately reveals that PCs 4 to 8 can be neglected as they only contribute <6% to the overall variability in the data. Therefore the focus lies on the first three PCs, whose contributions to the overall variability are 54.6%, 28.6% and 13.3%, respectively. The loadings plot demonstrates that the HBA, PSA and HBD variables correlate highly with the first PC, the number of rings and the molecular weight are correlated considerably, whereas the remaining three descriptors have low correlations with PC 1 (below 0.5). The second PC is dominated by the atomic composition and logP descriptors although the molecular weight variable's correlation with PC 2 still exceeds 0.5. This demonstrates that, in the first two PCs, all descriptors contribute highly, with the exception of the percentage of rotatable bonds variable, as also shown in the correlation circle in figure 4.7A.

Correlation circles are a visualisation of the extent to which each of the original variables contributes to a pair of PCs (projection from the unit sphere, see section 2.2.5.5). As mentioned above, figure 4.7A demonstrates that all variables, except for the percentage of rotatable bonds, contribute highly to the first two PCs. All observations from the previous paragraph can be confirmed here. In contrast to figure 4.6, which shows the contributions of each variable to each PC, figure 4.7 illustrates the contribution to pairs of PCs (PCs 1 and 3 in figure 4.7B, PCs 2 and 3 in figure 4.7C).

The placement of the cofactor HET labels in figure 4.5 is defined by each cofactor's value in the first two principal components of the scores matrix. The highest contributing individuals (*i.e.* cofactor HET identifiers) in these first two PCs are U10, MQ7, B12. This implies that these three molecules are the most unusual in their set of descriptors compared to the full cofactor data set. Their unusual properties have already been discussed in section 4.1.1. Thus, as expected, these cofactors are the outliers in the PCA. In spite of this, it has been decided to not exclude them from the analysis, since they are legitimate contributors to this data set.

With respect to finding intrinsic groupings in cofactors, the following visual clusters have been observed: the porphyrin-like molecules (HEM, HEA, HEC, B12, SRM, F43) as well as menaquinone-7 (MQ7) and ubiquinone-10 (U10) are all placed in the upper half of the plot, since these molecules are larger and on average more hydrophobic than the remaining cofactors. MQ7 and U10 are clearly the most hydrophobic cofactors, and within the large ones they are also the most flexible. Siroheme (SRM) and factor 430 (F43) are on the other end of the hydrophobicity spectrum. Heme A (HEA) is more hydrophobic and larger than heme C (HEC), but both have average flexibility within this group of large cofactors. Vitamin B12 (B12) is by far the largest of the cofactors and less hydrophobic than most of the molecules in the large cofactors.

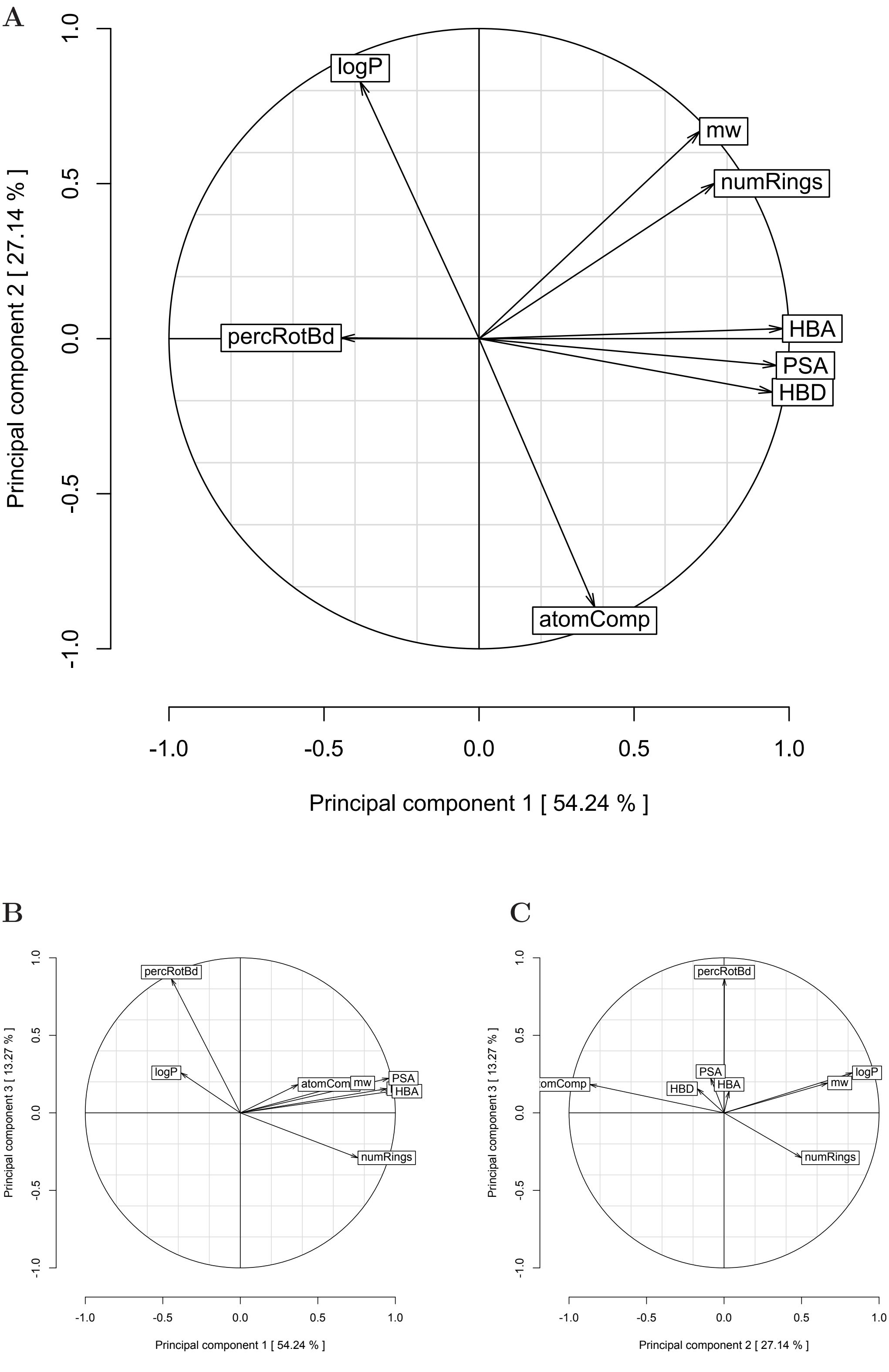


Figure 4.7: Correlation circles of the eight descriptor variables in the first three principal components: (A) PC 1 and 2, (B) PC 1 and 3 and (C) PC 2 and 3.

On the right-hand side of the plot, all the dinucleotides (NAD, NAP, FAD, COA, PGD, MGD and PTE) can be found. Of course, the molecules in this group have a similar atomic composition, which is rich in heteroatoms in comparison to heme and menaquinone. Among the cofactors, the dinucleotides are of medium to high molecular weight. They separate in the first principal component due to differences in hydrogen-bonding potential (HBA and HBD variables) and PSA.

The remaining cofactors in the bottom left corner of the PCA plot are smaller with lower hydrogen bonding potential, but they are much more varied as a group.

Overall, these results reveal the similarities between the various cofactors, as well as between the cofactors and the variables, and show that cofactors vary mostly in polarity and size. Although one starts to see intrinsic groupings in the PCA, it is advisable to apply clustering methods to the cofactor data in order to clarify the results and strengthen the confidence in theses groupings (see sections 4.1.5, 4.2 and 4.3.1).

4.1.4 PCA of metabolite descriptors: locating cofactors in metabolite space

In order to locate the cofactors in "metabolite space", another PCA has been performed on the metabolite data set (see section 2.2.1.2, grey points in figure 4.8). The exact same methodology has been used as in the PCA described in the previous sections. This analysis aims to determine whether cofactors are restricted to an easily distinguishable corner of metabolite space (*i.e.*, originate from a limited homogeneous set of chemical moieties). To answer this question visually, the cofactors are plotted as supplementary objects [1] (red circles) in figure 4.8. The biplot (analogous to figure 4.5) and the loadings plot (analogous to figure 4.6) are shown in figures A.3 and A.2 in the appendix, respectively.

Figure 4.8 shows that the cofactors (red circles) are composed of chemical moieties, which span the full range of biochemical small molecules, rather than being confined to one specific physicochemical group. This implies that cofactors are not basically different from other metabolites, as measured by the chosen

[1] Supplementary objects are not used to compute the principal components but can be plotted together with the other data points.

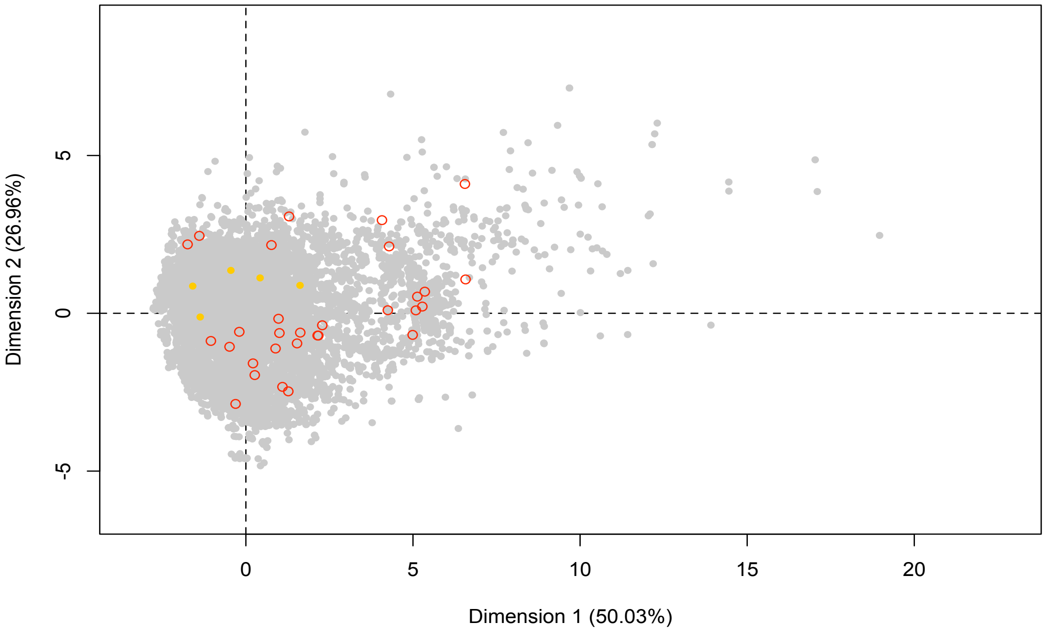


Figure 4.8: The first two principal components of metabolite space, showing the metabolites (grey) and the supplementary cofactor data (red). Example metabolites in the cofactor free space are shown in yellow.

descriptors, and cannot be easily distinguished. As expected, the correlation matrices (see figure A.1 in Appendix A) of the descriptors are similar for both data sets (cofactors and metabolites). The yellow data points in figure 4.8 are representatives of a cofactor-free space, which we grouped with medium-sized hydrophobic molecules, presumably because they are not capable of providing the required catalytic power. The yellow data points denote examples from this area, which include the cholane-based bile acid taurochenodeoxycholate (KEGG compound ID C05465), the indole alkaloids echitovenine (C11784) and strictosidine (C03470), the frog toxin epibatidine (C11690), as well as the leutokrine A4 (C00909). All of these molecules are largely hydrophobic and are medium to large in size. They do bind to proteins but lack the catalytic power that cofactors provide. However, overall one can see that cofactors have been drawn from the whole spectrum of biochemical metabolites and thus are constructed from many of the basic building blocks of life (see section 4.2.3).

4.1.5 Intrinsic groupings among cofactors: a 1D clustering

Clustering methods assign objects (here the cofactor HET groups) to several classes, according to their pairwise similarity. Figure 4.9 shows the results of a Ward's clustering (Ward Jr, 1963) of the cofactor descriptors. The figure depicts a hierarchical clustering tree representing these pairwise similarities. A penalty function (Kelley *et al.*, 1996) has been used to cut the tree at the best height in order to obtain the final clusters.

To quantify the overview of all the cofactors, it is worth exploring whether it is possible to automatically cluster cofactors into groups with similar properties. A hierarchical Ward's clustering of the cofactor descriptors is presented in figure 4.9B. The agglomerative coefficient of the data on the tree is 0.93. In order to find the best number of clusters, the tree is cut at the height, which is determined by a minimum penalty score (Kelley *et al.*, 1996, see figure 4.9A). This penalty score finds the number of clusters, which minimises the spread within the clusters while maximising the population density of the clusters. The minimum of this penalty function determines how many clusters should result from the dendrogram cut. Here, this method yields six clusters as the optimum. The first split in the tree separates clusters 1-4 from 5 and 6. This split is based on polarity and size, much like the driving force in principal component one (see figure 4.5). The molecules on the right side of the split (clusters 5 and 6) are more polar and larger than the remaining ones. The subsequent separation of cluster 4 from 1-3 separates the very hydrophobic large molecules from the others. This is also analogous to the second principal component of the cofactor PCA. Similarly, the split of cluster 3 separates the very flexible cofactors (third principal component), and the remaining two clusters split on size and ring content. All the six clusters can be drawn into figure 4.5 without overlap (see figure 4.10), except for clusters 1 and 3. The overlap between clusters 1 and 3 is not surprising, since the determining factor for this split is flexibility, which has been shown in this chapter to be

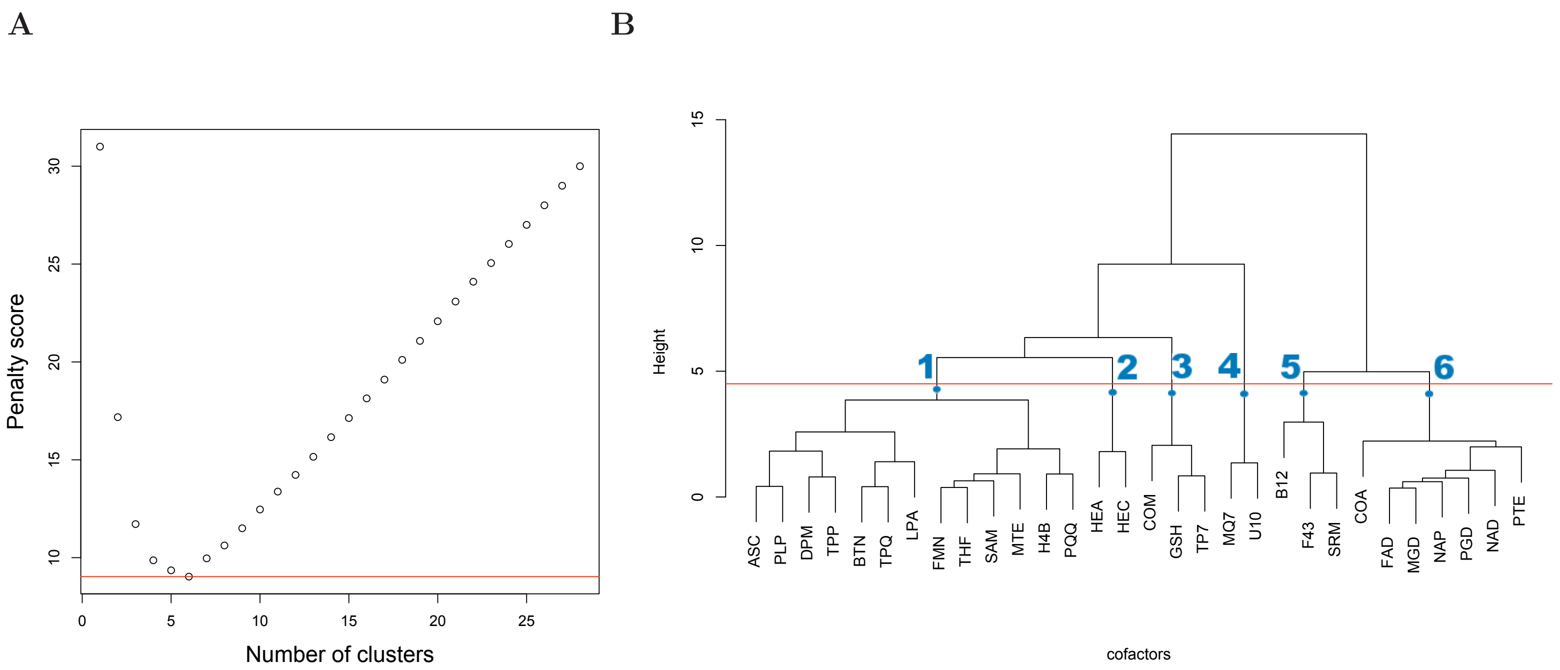


Figure 4.9: Hierarchical Ward's clustering of cofactor descriptors. A: The tree is best cut into six clusters according to the penalty score. The minimum penalty occurs at 6 clusters and is marked by the horizontal red line. B: The six clusters are: 1, small ring-containing cofactors; 2, large hydrophobic rigid cofactors (=hemes); 3, small flexible cofactors; 4, large hydrophobic flexible cofactors (large quinones); 5, non-hydrophobic large cofactors (the polar porphyrin-like cofactors) and 6, medium-size charged cofactors (=dinucleotide-like).

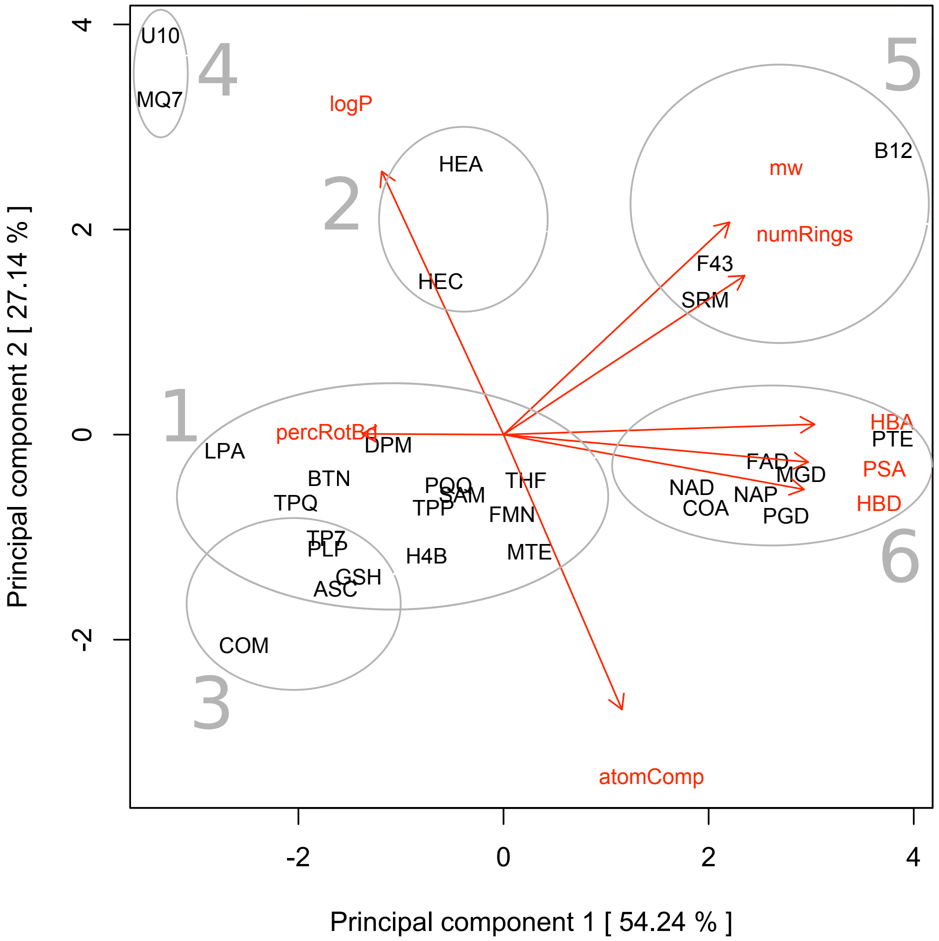


Figure 4.10: Biplot of the descriptor PCA with clusters from figure 4.9B. Only clusters 1 and 3 overlap.

the main contributor to the third principal component. Therefore, those two clusters split in the third dimension of the PCA. Cluster 1 comprises all the small (molecular mass $\leq$500 Da), ring-containing cofactor molecules. The first split inside this cluster is based on the percentage of rotatable bonds (35-55%, left; 23-45%, right), the number of rings (1-3 left, $\geq$3 right) and the number of HBDs ($\leq$4, left; $\geq$4, right). Cluster 2 contains HEA and HEC, whose similarities have been described previously (section 4.1.3.1). Glutathione (GSH), coenzyme B (TP7) and coenzyme M (COM) are in cluster 3. These molecules have an unusually high percentage of rotatable bonds (high level of flexibility) for cofactors and all of them contain a sulfur atom. Cluster 4 contains the two large, hydrophobic cofactors: MQ7 and U10 (coenzyme Q10), which are structurally very similar to each other and also clustered together in the PCA. Cluster 5 contains the porphyrin-based cofactors adenosylcobalamin (B12), F43 and SRM. They share the same basic scaffold as the hemes (HEC, HEA), but are less hydrophobic due to a lower double bond content, as well as due to carbolic acid and amine groups as side chains (SRM, F43). Once more, a structural similarity can be identified in this cluster. Cluster 6 also contains all the dinucleotides (NAD, NAP, FAD, PGD, MGD), the structurally very similar tungstopterin cofactor (PTE), and coenzyme A (COA). This cluster corresponds to the dinucleotide cluster in figure 4.5. These results indicate that the chosen descriptors partition the cofactors well into six groups according to their physicochemical properties. The structural similarity, which is shared among the cofactors within each cluster, supports this conclusion.

4.2 Analysis of two-dimensional cofactor structures

The one-dimensional descriptor variables have provided valuable insights. The aim of this section is to investigate if and how information about the two-dimensional structure of the molecules complements or confirms the above findings.

4.2.1 An overview of the functional portions in cofactors

First it is prudent to take a closer look at the functional portions and two-dimensional structures of all the cofactors in the data set. Table 1.1 lists the structures of all cofactors in the data set. The active portions are highlighted in magenta.

4.2.2 Intrinsic groupings among cofactors: a 2D clustering

The SMSD program (Rahman *et al.*, 2009) compares two molecules and assigns a similarity score based on their common substructure (see section 2.2.4). The comparison can be set to distinguish between single and double bonds ("bond-sensitive") or to not distinguish between them ("bond-insensitive"). SMSD has been used to generate two pairwise Tanimoto similarity matrices, one bond-sensitive and one bond-insensitive one. These matrices were then used to cluster the cofactor data set once again using Ward's criterion. Figure 4.11 shows the resulting clusterings (B and D) and penalty functions (A and C) for the bond-sensitive and bond-insensitive similarity matrix, respectively.

The agglomerative coefficient of the bond-sensitive clustering is 0.78, while the one of the bond-insensitive clustering is 0.83. Although both of those values are lower than the descriptor-based clustering, at least the bond-insensitive coefficient promises useful insights, as the data fits the clustering well.

4.2.2.1 Bond-insensitive SMSD clustering

The bond-insensitive clustering allows for an analysis of the intrinsic groupings among the cofactors based on their general chemical scaffold. Figure 4.11D shows the resulting dendrogram and the best (solid line) and second best (dashed line) dendrogram cut. Both lines are shown here since the penalty score function has two minima with very similar values, as can be seen in figure 4.11C. The solid line produces four clusters.

A similarity matrix has been generated based on a pairwise bond-insensitive comparison of all molecules in the data set to look for similarities in the ge-

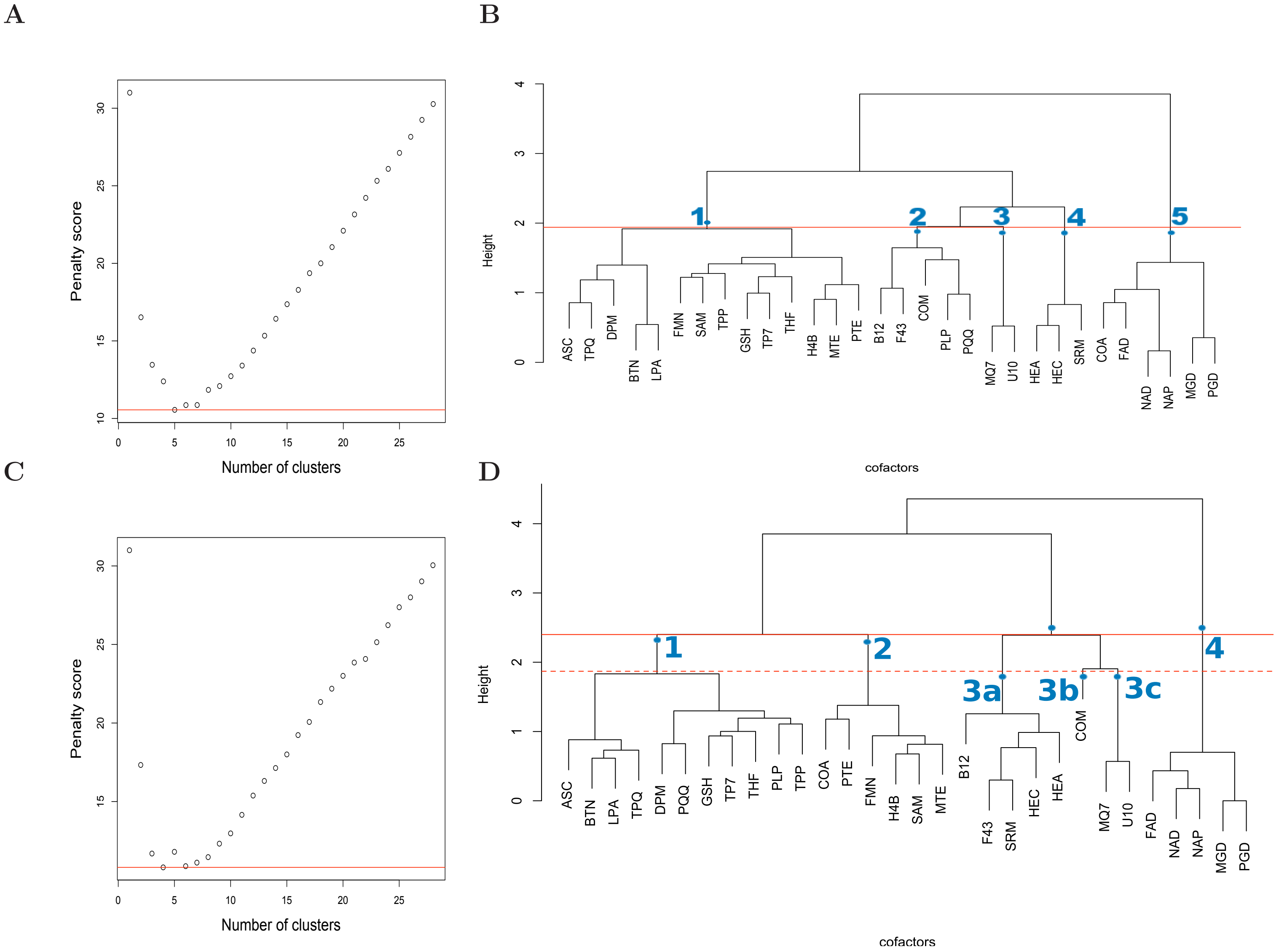


Figure 4.11: Bond-sensitive (B) and bond-insensitive (D) clustering of the two-dimensional cofactor structures, based on a Ward's clustering algorithm using a Tanimoto similarity matrix. Plot of the penalty function for the dendrogram cut to produce any number of clusters for bond-sensitive (A) and bond-insensitive (C) clustering. The minimum penalty occurs at 5 and 4 clusters respectively and is marked by the horizontal red line.

neral molecule scaffold. Distinguishing double from single bonds would result in a higher differentiation of function and hydrophobicity of the molecules, but information about the generality of the scaffold would be lost. This scaffold information, however, is important from a biosynthetic point of view, as similar scaffolds might be biosynthesised by the same enzymes. The agglomerative coefficient of the data on this tree is 0.83, which shows that the tree fits the data well, although not as well as when the physicochemical descriptors are used (agglomerative coefficient 0.93).

The penalty function, which decides at which height the tree is best cut, has a minimum at four clusters. However, it is worth noting that the split between clusters 1 and 2 and the first split inside cluster 3 are at almost the same height. The penalty function has another local minimum at seven clusters, which would split cluster 3 into three clusters (dashed line). In comparison to the descriptor-based clustering (figure 4.9), the same separation by size is observed, but instead of the separation by polarity, a better structural scaffold resolution is achieved. Cluster 3 contains all the very large cofactors. All five porphyrin-based cofactors are now grouped together (subcluster 3.1), as they share the same principal molecular scaffold, although they differ in polarity. The other two very large cofactors (subcluster 3.3) have a quinone portion in common with a long carbon side chain. If cut at the solid line, they are located in the same cluster as the porphyrins, but if the dashed line is used for the dendrogram cut, they separate. Coenzyme M (subcluster 3.2) appears to be misclassified by the solid line, but would be a singleton when the tree was cut at the dashed line. Since a bond-insensitive clustering was used, flexibility, which classified COM with TP7 and GSH in figure 4.9, cannot contribute extensively here. Cluster 2 in figure 4.11D, the mononucleotides, was not detected in the descriptor-based clustering (figure 4.9), and its members were allocated to clusters 1 (small, ring-containing) and 6 (dinucleotide-like) there. Due to these molecules' similarities to the members of clusters 1 and 6 in physicochemical properties, the assignment was correct from a descriptor variable point of view. However, the use of two-dimensional structural data allows to make this distinction here (figure 4.9).

The dinucleotides cluster 4 in figure 4.11D contains all the "real" dinucleotides and corresponds to the inner five molecules of cluster 6 in the descriptor-based tree. All of the real dinucleotides have an adenosine or guanosine portion, which accounts for their high pairwise structural similarity. The two molecules, which are classified as “dinucleotide-like” in the descriptor-based but not in the structural clustering, are COA and PTE. Although COA has a full adenosine portion, the other half of the molecule is structurally very different from a nucleotide, as it does not have any pyrimidine- or purine-based ring structures. Therefore, its structure resembles a mononucleotide rather than a dinucleotide and thus it correctly appears in the mononucleotide cluster in figure 4.11D. The tungsten cofactor PTE consists of two identical halves, each of which resembles a mononucleotide. Since the connection between those two halves is realised by a tungsten atom surrounded by the four sulfur atoms at one end of the molecule and a magnesium atom bound by the two phosphate portions at the other, the halves are not connected in the same way as the other dinucleotides (phosphodiester bond). PTE is therefore correctly clustered with the mononucleotides, rather than the dinucleotides, on a structural similarity basis. Cluster 1 in figure 4.11D corresponds mostly to cluster 1 in the descriptor-based clustering excluding the mononucleotide-like molecules: it contains the low-molecular-weight cofactors with a small number of rings. In this structure-based clustering, however, GSH and TP7 (both in cluster 3 in figure 4.9) are correctly grouped with the other small molecules, most likely again due to the lack of flexibility information. In summary it can be said that the structural clustering reveals a better classification of the cofactors on their basic molecular scaffold. The relevance of the scaffold to cofactor evolution from building blocks of life is even more clearly visible.

4.2.2.2 Bond-sensitive SMSD clustering

To complete the intrinsic grouping analysis, the bond-sensitive analysis has also been performed. This molecular comparison is more sensitive to the way the molecules are stored with respect to their electronic configuration. Compared to the bond-insensitive comparison, this method provides a “higher resolution” on the bond similarity, but loses the generality of the bond-insensitive approach.

Figure 4.11B shows the dendrogram that results from the Ward’s clustering of the bond-sensitive similarity matrix. The tree is best cut into 5 clusters, as defined by the penalty score function (see figure 4.11A: minimum at 5 clusters). Cluster 1 in figure 4.11B comprises all the small cofactors and mononucleotides and thus corresponds to the union of clusters 1 and 2 in the bond-insensitive clustering, with the exception of PLP, PQQ and COA. PLP and PQQ have a pyridine substructure in common, which groups those molecules closely together. They appear in cluster 2 of the bond-sensitive clustering, together with COM, B12 and F43. The similarities between F43 and B12 have been discussed earlier. It is interesting to note that, firstly, F43 and B12 are not in the same cluster as the other hemes HEA and HEC, like in the descriptor-based clustering (see figure 4.9B) and that, secondly, they here are grouped together with COM, like in the bond-insensitive clustering. The reason for the first observation is most likely, when comparing the overall similarity between F43/B12 and HEA/HEC/SRM, that F43 and B12 have too many different additions at various locations in them meta ring system, such that the overall substructure match is smaller than in the bond-insensitive clustering because the additions modify the double-bound content of the meta ring system. The second observation, that they are grouped with COM in both two-dimensional structure-based clusterings, can be explained by the general uniqueness of these molecules. B12 is the most prominent "outlier" in the data (see figures 4.1 and 4.5) but from a structural point of view it has significant common substructure with F43. COM is structurally so different from any of the other cofactor structures that the clustering algorithm has difficulties placing it in the two structure-based dendrograms. As COM is a very small molecule, the best common substructure match occurs with the phosphate portion of PLP, which is the smallest phosphate-containing cofactor in the data set. The lack of commonality of COM, B12 and F43 most likely defines cluster 2. This hypothesis is supported by the height at which F43 and B12, but especially COM is placed, as the height of a leaf in a dendrogram indicates how confident the placement decision is.

Clusters 3 and 4 in figure 4.11B contain the large quinones and hemes respectively. These molecules' similarities have been discussed at various occasions and are easy to understand when comparing their structures in table 3.1.

Lastly, cluster 5 is identical with cluster 4 in the bond-insensitive clustering, with the addition of COA. It is interesting to note, that the left subcluster contains

COA, FAD, NAD and NAP who all share a adenosine substructure, whereas the right subcluster comprises MGD and PGD who share a Guanosine substructure.

The bond-sensitive clustering has been performed for completeness but did not reveal insights that exceed the ones from the descriptor-based and the bond-insensitive clustering although it as further confirmed some of the groupings found before.

4.2.3 Building blocks of life

The results above demonstrate that the cofactors consist of often similar chemical scaffolds and functional groups. This section discusses how some of these portions are common metabolites in the cell themselves. To follow the structural similarities mentioned, please refer to table 1.1 and figures 4.12 and 4.13.

Many cofactors have full nucleotide moieties (see figure 4.13A), or parts thereof (see figure 4.13D and E) incorporated, which gives them the ability to be bound to an RNA molecule (see figure 4.12D and E). It has been hypothesised (Muller, 2006) that this could be a remainder from cofactor-using ribozymes in an RNA world. However, RNA residues are not the only building block of life that can be found in cofactors. Many cofactors also utilise amino or fatty acid building blocks. For example, GSH is made up of the three amino acids glutamate, cysteine and glycine, where the glutamate is attached *via* its γ-carboxyl group (see figure 4.13C). The MIO cofactor spontaneously self-catalyses (Christianson *et al.*, 2007b) from the Ala-Ser-Gly sequence in the protein sequence to build the cofactor, which possesses catalytic properties different from those of the three underlying amino acids (see figure 4.13B). Furthermore, coenzyme A's active portion is built from the two amino acid derivates decarboxylated cysteine and beta-alanine as well as the vitamin pantothenic acid (see figure 4.13D). Finally, it is worth noting that the lipoic acid (LPA) prosthetic group is partly composed of the fatty acid octanoic acid (Marquet *et al.*, 2001).

It has long been known that the sulfur-containing amino acid cysteine has unique catalytic properties, especially for free radical chemistry (see e.g. Plaga *et al.* 2000). Many cofactors also have one or more sulfur atoms. Examples include LPA, SAM, biotin (BTN) and thiamine diphosphate (TPP), as well as

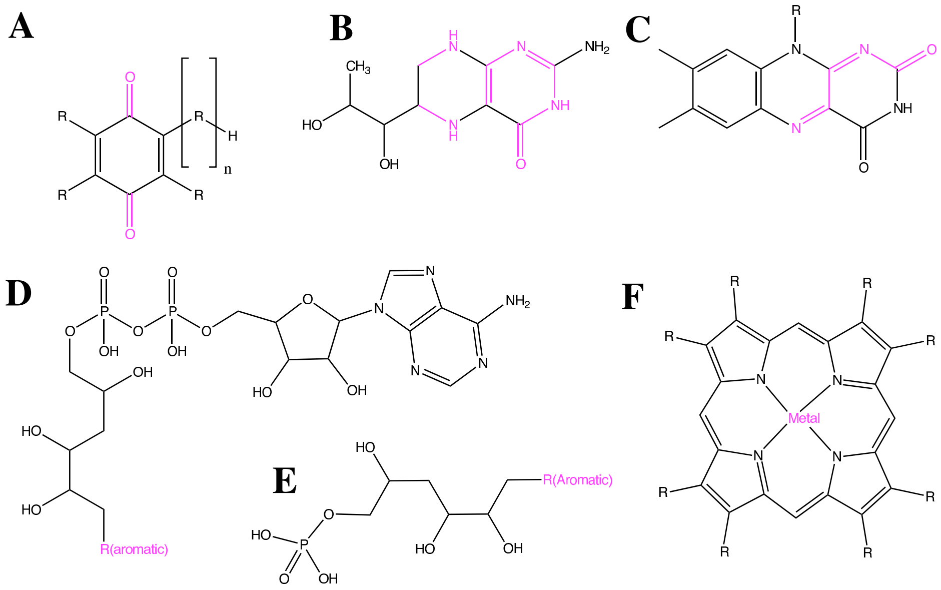


Figure 4.12: Recurring structural motifs in the organic cofactor data set. Functional portions are shown in magenta. A: quinone-containing like MQ7, U10, B: pterin-based structure of biopterin, C: flavin, as occurring in FAD and FMN, D: a dinucleotide like FAD, E: mononucleotides like FMN and F: porphyrin-based scaffolds like HEM.

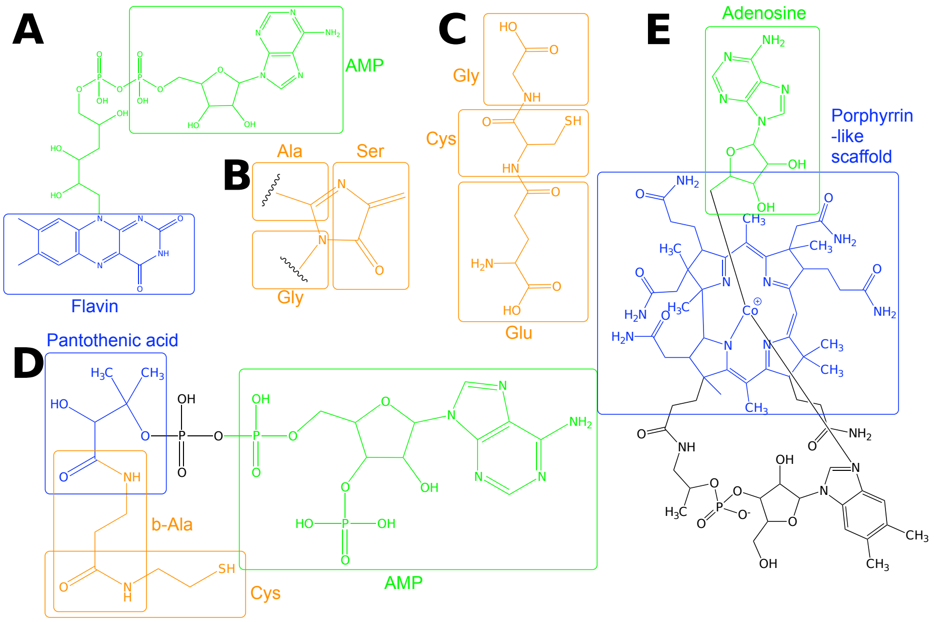


Figure 4.13: Examples of cofactors made from building blocks. The nucleotide building blocks are shown in green, the amino acid ones in orange and the cofactor-specific scaffolds in blue. A: FAD, B: MIO, C: GSH, D: COA, E: B12.

the aforementioned cofactors COA and GSH. In LPA, SAM, GSH and COA, these sulfur atoms are directly involved in catalysis, which suggests that the cell relies on these cofactors for similar reasons to its utilisation of cysteine. In TPP and BTN, the sulfur atoms are not directly involved in catalysis and are part of heterocyclic components where the catalytically active atom is a carbon and a nitrogen atom, respectively. The sulfur atom in thiamine is located adjacent to the catalytically active C_4 atom, which becomes a carbanion in the activation step of TPP's catalytic mechanism (Tittmann *et al.*, 1998). One can therefore safely assume that the sulfur atom's electron-withdrawing properties are crucial in this cofactor. The functional role of the sulfur in BTN remains unclear.

Another very common motif among the cofactors is the pteridine scaffold (see figure 4.12B), which appears in the folic acid compounds, namely THF, MHF, PTE and H4B, as well as in the flavins (FAD and FMN, see figures 4.12C and 4.13A). Please note the large common substructure between the pterins and the flavins. Although pteridine is not one of the basic building blocks of life, it is directly synthesised from the RNA building block GTP (Holliday *et al.*, 2007b).

The quinones and the porphyrin-based scaffolds (see figures 4.12A and F, respectively) seem to be specific to cofactors. Pyrroloquinoline quinone (PQQ) and TPQ are water soluble, whereas MQ7 and U10 are fat soluble. The porphyrin scaffold appears with three different catalytic metal ions (iron, nickel and cobalt) and seems to be adjusted to its task by varying side chains and double bond content as well as the type of metal in the centre. Please note that the pteridine-based cofactors, as well as the quinones and porphyrins, seem to be mainly involved in electron transfer and redox chemistry functions, which are almost never catalysed by amino acid residues (see figure 4.16).

Some cofactors, such as B12, SAM or COA, combine different building blocks. Adenosylcobalamin (B12), as the name suggests, combines a base (RNA building block) with a (porphyrin-like) corrin system (see figure 4.13E), SAM combines its adenosine portion with the amino acid methionine (a protein building block), and COA combines it with pantothenic acid, which is in turn made up of the vitamin pantoinate, beta-alanine, and cysteamine, a decarboxylated amino acid. Further, PLP (pyridoxal-5'-phosphate) is based on a pyridine and a phosphate group. From this, it can be seen that cofactors are primarily constructed from the existing building blocks of life. Most of those building blocks have other essential functions in the cell (nucleotides, amino acids, fatty acids), whereas a few seem to have exclusively evolved to be cofactor-specific, complementing the cell's catalytic toolkit.

4.3 Functional analysis of cofactors

While the one- and two-dimensional properties have revealed those properties of the organic cofactors that do not require any knowledge about their role in enzymes, the function of these molecules is analysed in this section in order to complement the knowledge gained so far. Chapter 5 is devoted to the analysis of three-dimensional conformations.

There are two levels of detail, at which the function of a cofactor-dependant enzyme reaction can be analysed: the level of the overall reaction that is catalysed by the enzyme and the mechanistic level that reveals how the enzyme achieves the overall transformation.

4.3.1 Manual classification based on overall reaction

The overall reaction of an enzyme is hierarchically classified by the enzyme commission (E.C., NC-IUBMB & Webb, 1992). The E.C. assigns a number to each overall enzyme reaction. This number has 4 levels of hierarchy. The first and most general level classifies the enzyme into one of six possible enzyme classes: the oxidoreductases (1), the transferases (2), the hydrolases (3), the lyases (4), the isomerases (5) and the ligases (6). The numbers on the second and third levels describe the chemistry that occurs in increasing detail. The last number is a serial number and usually distinguishes substrate specificity of otherwise similar reactions.

From the above clustering results, it is clear that the physicochemical descriptors and the structural similarities yield reasonably good distinctions between the different cofactor groups. However, they do not correspond to a functional classification. Therefore, a knowledge-based classification of cofactors is proposed, which is determined by each molecule's overall effect in enzyme reactions. This is shown in figure 4.14. Seven distinct, but not mutually exclusive functional roles are identified. The functional roles are based on the overall chemical changes in the substrate(s) of the enzyme reactions, which are co-catalysed by each cofactor. These are:

1. Moiety transfer: transfer of chemical moieties, such as functional groups. It should be noted that this excludes the moiety transfers involved in redox chemistry, that is, hydrogen species (H^+, H^- and $H\cdot$) and electron transfer. For example, BTN transfers CO_2 moieties between substrates and/or a substrate and the solvent. This class of function is shown in the red circle in figure 4.14.

2. Redox: hydrogen species and electron transfer (including radical formation), as well as general redox chemistry (the blue ellipse in figure 4.14).

3. Mobility: the conformational flexibility of the cofactor is used to pass the substrate/intermediate around between different active sites (orange ellipse in figure 4.14).

4. Activation: photochemical activation of aromatic systems, which then act as a reaction initiator (black ellipse in figure 4.14).

5. Bond cleavage/formation: cleavage or formation of a bond between heavy atoms (green ellipse in figure 4.14).

6. Polymerisation: DPM is the only representative of this role, which is interesting, as it is uniquely associated with the biocatalysis of the porphyrin ring, a fundamental building block of other cofactors. Polymerisation (cyan ellipse in figure 4.14) is a special case of bond formation.

7. Rearrangements: cofactors effecting intramolecular rearrangements of atoms in the substrate, shown in the purple ellipse in figure 4.14.

This classification is based on currently published scientific knowledge as well as the functional detail in the MACiE database (Holliday *et al.*, 2007a). In many ways it mirrors part of the primary E.C. classification used to describe enzyme

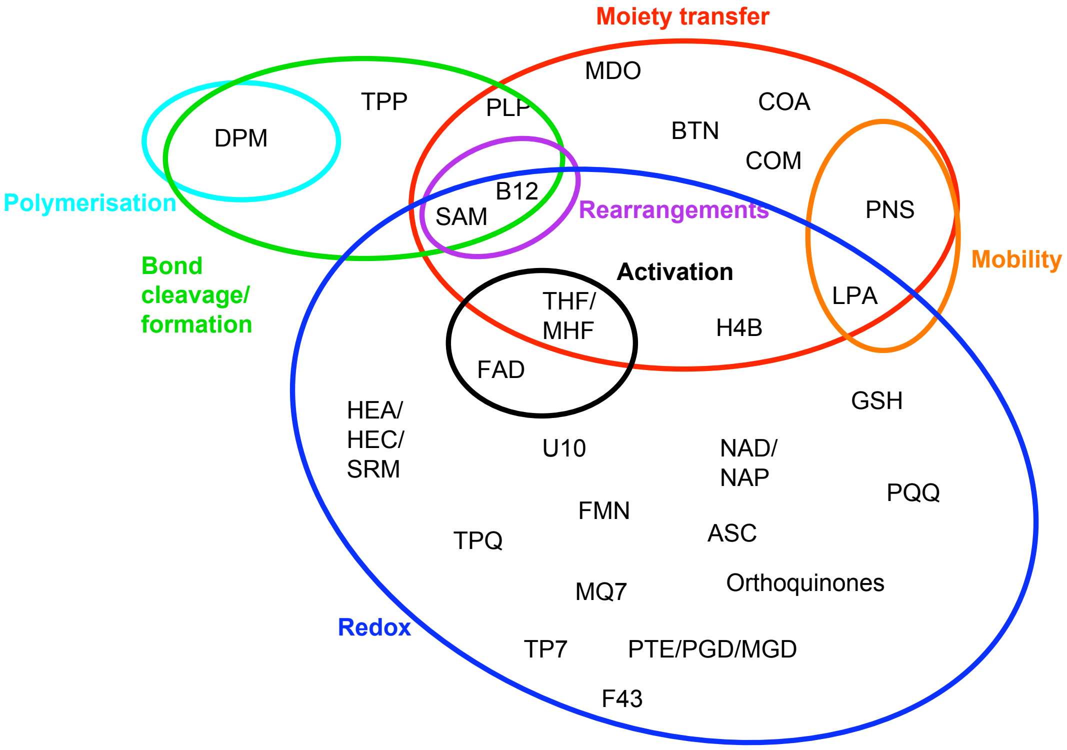


Figure 4.14: Functional classification (Venn diagram) of cofactors, based on their effect in the overall reactions they take part in.

function. It is entirely possible that future research will reveal both new categories and new memberships in existing categories. Whilst not immediately obvious from the above figure, except for the case between B12 and heme, it is interesting to note that cofactors can have a very high structural similarity, but rather different physicochemical properties and overall effects in their host enzymes. For example, FAD, folic acid and biopterin have a large common substructure, but their (known) effect profile is rather different: while FAD mostly acts as a redox cofactor, folic acid mainly transfers C_1 groups and biopterin can transfer hydroxyl groups as well as engage in redox activities. The porphyrin-like cofactors are another demonstration of this point: while the hemes (HEC, HEA, SRM) act as redox cofactors, cobalamin's (B12) overall effect may further be a molecular rearrangement in the substrate or methyl group transfer.

In the terminology used by enzymologists and biochemists, the term "cofactors" mostly refers to the moiety transfer and redox cofactors defined above. It is tempting to assume that the moiety transfer cofactors are coenzymes as defined in chapter 1 (*i.e.* they bind to the enzyme at the beginning of each catalytic cycle, much like the substrates), whereas the redox cofactors are prosthetic groups (*i.e.* they are inserted into the enzyme once and then stay there). However this is not always the case. B12, for example, is always a prosthetic group, while it may catalyse both moiety transfer (methyl groups) and redox chemistry. SAM, on the other hand, has the exact same overall function profile as B12, but is a coenzyme. Further, the MIO cofactor is a prosthetic group (polypeptide-derived cofactor) but acts as a NH_2-transfer cofactor. There are many more examples like this. Please refer to the database for coenzyme/prosthetic group information.

4.3.2 Participation of cofactors in the six enzyme classes

Next, the coverage of enzyme space (as defined by the E.C. classification) by organic cofactors is examined. Figure 4.15 shows the total number of active entries in the E.C. database, which CoFactor annotated as being organic cofactor dependant (shown in dark grey), and the number of entries, which are not annotated as organic cofactor dependant (i.e., are considered to be organic cofactor independant), shown in light grey. These data are plotted for the six enzyme classes, respectively. The percentage of enzymes that use a cofactor within each of the six classes is shown below each bar.

Figure 4.15 demonstrates that E.C. class 1 (the oxidoreductases) is most dependant on organic cofactors, since more than 80% of them use at least one. Some E.C. numbers utilise several different cofactors in several different enzymes, such as E.C. 1.11.1.10, in which there are three "types," one of which is cofactor independant (MACiE entry M0248), one of which uses heme (MACiE entry M0250), and one of which uses the inorganic vanadate cofactor (MACiE entry M0014). In the other classes, the percentage of cofactor-using enzymes is around 30%, with the exception of the hydrolases (E.C. class 3), which rarely employ organic cofactors (4%).

A plausible explanation for the low frequency of organic cofactor usage in the hydrolases is, in part, the relative simplicity of these mechanisms: 79% of

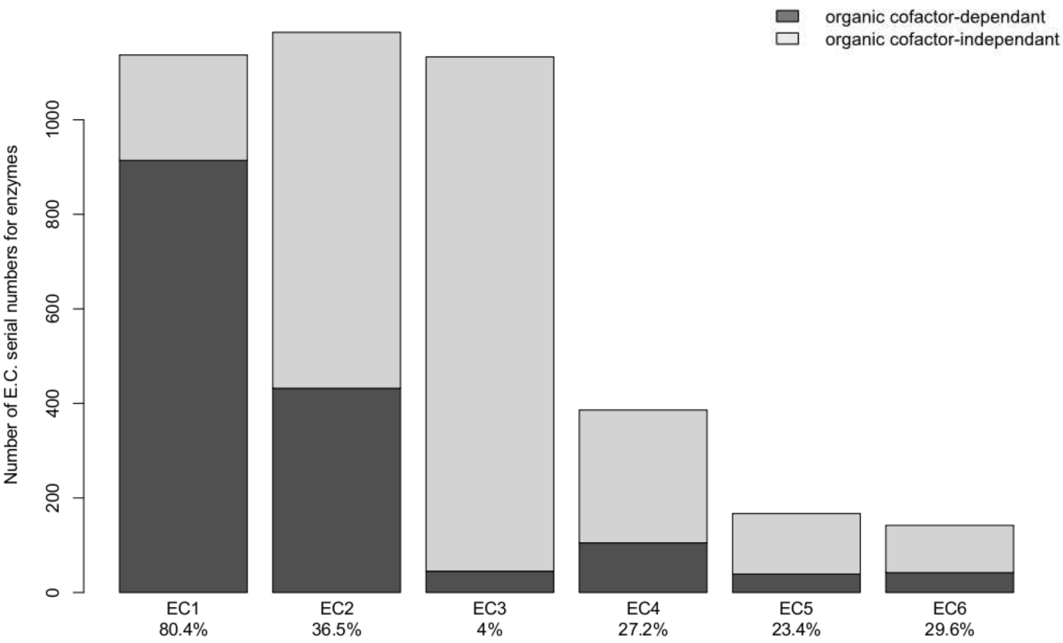


Figure 4.15: Dependence on organic cofactors in the six E.C. classes of the Enzyme Commission database: organic cofactor independant (light grey) and organic cofactor dependant ones (dark grey, enzyme data set). The percentage of organic cofactor dependant enzymes in each class is shown below the labels.

hydrolase mechanisms in the MACiE database consist of four or less steps. Furthermore, many metal ions, such as magnesium, can act as Lewis acids, activating water for an initial nucleophilic attack. These facts make inorganic cofactors more prevalent in this E.C. class (Andreini *et al.*, 2009; Holliday *et al.*, 2009).

The disproportional representation of organic cofactor dependence in the oxidoreductases can be explained by analysing the functional roles performed by organic cofactors (see the section 4.3.3) and the type of reactions being performed in this class of enzyme. For example, in order to reduce an alcohol to an aldehyde, two protons and two electrons need to be added. This is usually done – especially in those cases where NAD(P)H or FAD are involved – by the addition of a hydride ion (a hydrogen atom with two electrons associated) and a proton;

thus, the ability of certain organic cofactors to shuttle these hydride ions between molecules is vital to many of the oxidoreductase reactions. Approximately 97% of the organic-cofactor-dependant reactions in the oxidoreductase class of enzyme in the CoFactor (Fischer *et al.*, 2010b) database utilise at least one of those cofactors classified in the redox portion of figure 4.14. In the transferase (E.C. 2), lyase (E.C. 4), isomerase (E.C. 5) and ligase (E.C. 6) classes, 30±7% of these enzymes depend on organic cofactors. Of these, approximately 81% use at least one of the cofactors from the moiety transfer cofactor class in figure 4.14.

4.3.3 Molecular functions of organic cofactors in comparison to metal cofactors and amino acids

In order to investigate the overrepresentation of cofactor usage in the oxidoreductases, the roles that cofactors play in the detailed chemical stepwise mechanism (rather than the overview reaction used above) were analysed. Here, the functional role ontology used in the MACiE database was used, which describes nine distinct functional roles of chemical species in enzyme reaction mechanisms (Holliday *et al.*, 2007a, 2009). These terms can then be used to describe the roles of amino acid residues, as well as organic and inorganic cofactors. MACiE describes the enzyme's reaction mechanism in a stepwise manner and takes its information about each enzyme in the database from the literature. Using MACiE, the functional roles of both the amino acid residues and cofactors have been analysed in order to see where cofactors extend enzyme functionality and where they complement it.

Figure 4.16 was kindly provided by Dr. Gemma Holliday and generated from the MACiE database. It shows to what extent the three different types of catalytic entities (amino acids, organic cofactors and inorganic cofactors) are involved in the nine categories of mechanistic function defined by Holliday *et al.* (2009). Please note that complex cofactors with metal ions as well as organic portions appear in both categories: organic and inorganic cofactors.

The nine categories are:

1. Electrostatic stabilisers: stabilise a reaction intermediate mainly through electrostatic interactions

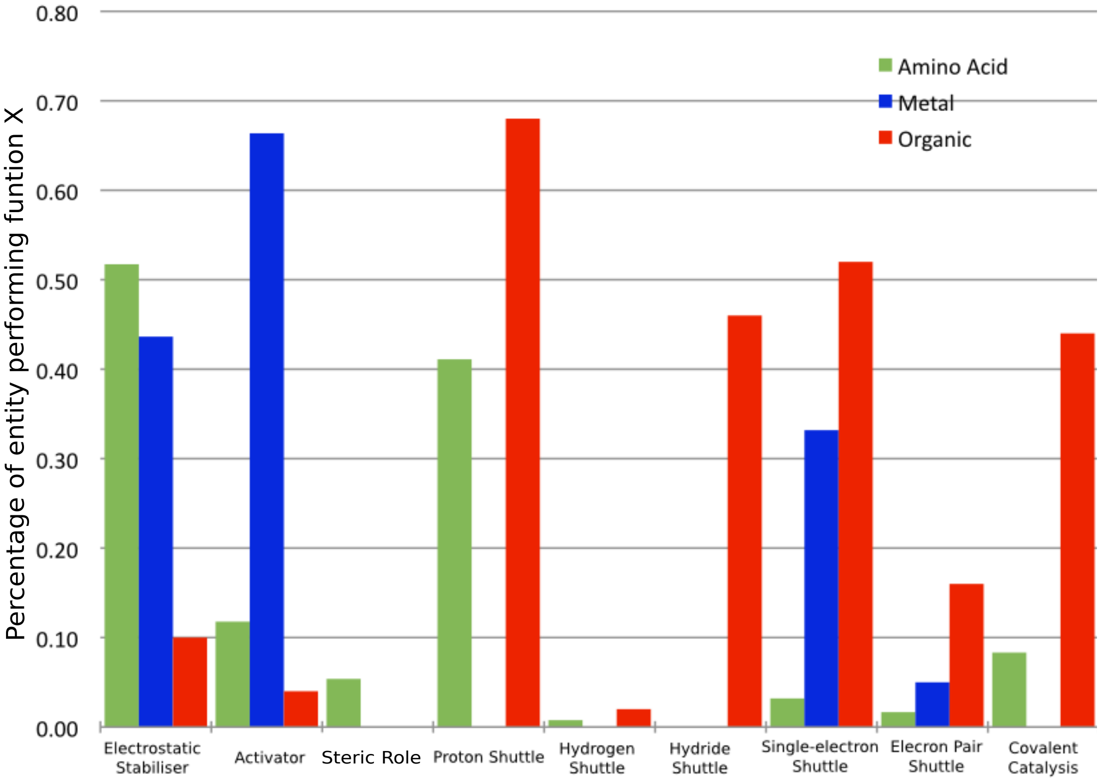


Figure 4.16: Figure showing the relative contributions of the three catalytic entities: standard amino acid residues (green), organic cofactors (red) and inorganic cofactors (blue) to the nine categories of functional role assigned in MACiE. There are a total of 1038 amino acid residues, 203 inorganic cofactors and 54 organic cofactors, and the height of the bars is normalised against the total number of entities annotated in MACiE. It is worth noting that any catalytic entity may be annotated with more than one of the nine functions (and therefore their normalised sum can exceed 1). For a more detailed description of these categories, see Holliday *et al.* (2009).

2. Activators: enable the reaction to occur (e.g. by modulating the pK_a or the redox potential)
3. Steric role: residues/cofactors direct the stereospecificity of an enzyme reaction (e.g. by steric hindrance)
4. Proton shuttle: denotes proton donators, acceptors and shuttles
5. Hydrogen shuttle: denotes hydrogen donators, acceptors and shuttles
6. Hydride shuttle: denotes hydride donators, acceptors and shuttles
7. Single electron shuttle: denotes single electron donators, acceptors and shuttles
8. Electron pair shuttle: denotes electron pair donators, acceptors and shuttles
9. Covalent catalysis: comprises cases where a proven covalent intermediate between the enzyme backbone and an intermediate was formed

From figure 4.16 it is immediately clear that organic cofactors are intimately involved in all aspects of catalysis. There are certain functional roles that are performed by all three catalytic entities, although to greatly differing extents. Other functions, however, are primarily performed by organic cofactors, like proton, electron (both singly and in pairs) and hydride transfers, and even in covalent catalysis. The amino acid residues are more critically involved in the formation of the local environment and in activating or stabilising the reactive intermediates. Even though organic cofactors are capable of this, they do it to a much lesser extent. It can further be seen that metal cofactors, like the amino acid residues, provide the appropriate local environment for a reaction to occur through their strong electrostatic interactions. Their electronic configurations additionally allow them to act in difficult redox (*i.e.* electron transfer) reactions.

While the qualitative profile of functional roles is similar between amino acid residues and organic cofactors, it is interesting to note that there is one function, namely the hydride (H^-) group transfer, which is only performed by organic cofactors in this data set. This demonstrates that, although many of the functions of organic cofactors may be also performed by amino acid residues, organic cofactors are crucial for catalysis in most oxidoreductases (see section 4.3.2). These E.C. class 1 reactions, where the majority of the hydride transfers occur, are only possible when organic cofactors are added to the enzymatic toolkit, as this data set suggests that hydride transfer is never performed by either amino acid residues or metal cofactors, but is instead unique to organic cofactors.

4.4 Discussion

Based on a thorough analysis of the scientific literature complemented by a statistical analysis of the physicochemical properties, chemical structures and overall functions of organic enzyme cofactors, this work has confirmed commonly accepted assumptions and, more importantly, offers a new way of considering and classifying these molecules. Whilst cofactors are built from across the pool of metabolites available to a biological entity, they tend to be a little larger – often being composed of two or more constituent building blocks – and considerably more polar. This need for increased polarity reflects their role as catalysts. Organic cofactors also add some missing and vital functionality to the enzyme's catalytic toolkit in the form of the hydride shuttling function, without which the redox chemistry of the oxidoreductase enzymes (coincidentally one of the largest classes of enzyme currently defined) could not occur efficiently.

Based on chemical structure, there are four different groups of cofactors (large cofactors, dinucleotides, mononucleotides and a somewhat broader group of other smaller cofactors). The first three groups are each composed of smaller constituent parts (which are often common metabolites in the cell, e.g., adenine), and the molecules within each of these three groups share large structural motifs, whereas the smaller cofactors (e.g., BTN) show little or no evidence of this. In contrast, the physicochemical properties split the cofactors into six groups, most of which have common structural motifs. This gives further support for the idea that many cofactors have evolved from smaller, more common building blocks. Organic cofactors are important across all areas of the chemical space of enzyme reactions, although obviously more so in the case of the oxidoreductase enzymes, and less so in the hydrolase class. Thus, it is unsurprising to see that the majority of cofactors identified in this study perform roles in either redox chemistry or moiety transfer. Only two cofactors are unassociated with either of these classes (DPM and TPP), both of which are involved in bond formation or cleavage of heavy atoms. However, whilst the structure of two cofactors can be rather similar, their functional profile can be very different (such as heme and B12). This raises interesting questions regarding the evolution of cofactors, including which functions came first, and which are the oldest cofactors. Are there evolutionary events that can be used to explain the need for a cofactor-dependant enzyme, and does this track with their evolutionary history as seen, e.g., for the co-evolution of alcohol dehydrogenase with the fleshy fruits of angiosperms and fermenting yeasts (Ashburner, 1998)? Which cofactors evolved from basic metabolites and which ones used another existing cofactor as a blue print? It seems likely that cofactors appeared very early on in evolution, since the existence of a cell without any of the cofactors analysed in this work seems unlikely. Further, several scientists have postulated a crucial role of cofactors in an RNA world (Jadhav & Yarus, 2002).

The catalytic range of cofactors covers all of the catalytic roles of amino acids observed today, which perhaps fits with a model of early evolution in which the full chemistry of life could be achieved using only ribozymes and cofactors. However, this is pure speculation, and it is clear that there is still much to be done in understanding how these small molecules add to the catalytic toolkit of enzymes and why the same molecule can perform very different functions in the cell. A quantitative analysis of the enzymes and domains that employ each cofactor and the cofactors' role in the evolution of life are intriguing subjects for further work in this area.

Chapter 5

The three-dimensional properties of organic cofactors

After the analysis of the one-dimensional physicochemical descriptor variables, the two-dimensional chemical structures and the functions of the organic enzyme cofactors in the previous chapter, this chapter concentrates on the analysis of the three-dimensional conformations and the solvent accessibility of organic cofactors in proteins.

First, the necessary background information will be provided, followed by the presentation of the results.

5.1 Background

The conformations that a small molecule adopts when bound to a protein has been a topic of interest for many researchers, especially in the context of pharmaceutical research (Davis *et al.*, 2008). Once a protein has been identified as a drug target, the next step is to find molecules that bind to that protein and thus may modify its activity in the patient's cells. Often drugs are designed to be inhibitors of an enzyme by either binding to the active site (substrate or cofactor binding site), or to an allosteric regulatory site, with a higher affinity then the naturally occurring ligands. Therefore, the drug molecule blocks the binding site and inhibits the reaction.

The analysis of ligand conformation can help answer important questions not only in pharmaceutical but also in basic research. Stockwell & Thornton (2006) conducted a detailed study of the three common protein ligands ATP, NAD and FAD. The authors found that all three ligands bind to proteins in varying conformations, although the variability of a ligand's conformation within a family of homologous enzymes is low in most cases. The authors further compare the observed ligand conformations from the protein crystal structures to the space of theoretically possible conformations and found that ligands tend to be in a more extended state when bound to the protein.

The conformation of a ligand determines its overall shape, which is one of the key factors in molecular recognition. Kahraman *et al.* (2007) have published a study in which they analysed the shape variation in protein binding pockets and ligands. The authors found astonishing differences in the shape and size of binding pockets that bind the same ligand. They reported that binding pockets are on average three times the size of the ligand. Several studies (e.g. Saranya & Selvaraj, 2009) have analysed the flexibility of active sites in several families of enzymes, and Weng *et al.* (2011) have recently published an analysis framework to investigate this issue on a larger scale from a perspective unbiased by manually selected protein families. Even artificial proteins have been found to bind a ligand using different binding modes (Simmons *et al.*, 2010). Given the active site flexibility and the amount of space that a ligand may therefore occupy in an enzyme's binding pocket, possibly in different binding modes, the question arises about what this implies for the conformational variability of the cofactors.

Based on the number and type of rotatable bonds, a small molecule can in theory adopt a large number of different conformations in three-dimensional space, which rises exponentially with the number of rotatable bonds (Vainio & Johnson, 2007) for any given level of detail (angle increment). In spite of this large number of theoretical conformations, the size and shape of the binding pocket, as well as steric hindrance among the ligand's own other binding molecules' atoms may impose restrictions on the number of feasible conformations in practise. In the case of the organic enzyme cofactors, which are a subset of protein ligands, the functional role they play in catalysis may further restrict functional conformations. To the author's best knowledge, no study has been published that systematically investigated this question. One of the further questions addressed in this chapter is, whether the findings about the conformational variability of ligands in the study by Stockwell & Thornton (2006) may be extended from the small data set of the three ligands ATP, NAD and FAD to all cofactors, for which sufficient structural data have been published.

Investigating the three-dimensional conformations of protein-bound small molecules crucially depends on the quality of the ligand models in the crystal structures of the proteins containing these ligands. The quality of ligand structure determination in protein crystals is often more error-prone than the quality of the protein structures (Davis *et al.*, 2008). Thus it is essential to assess the quality of the crystal structures for conformational analysis.

There are several methods for quality assessment of protein structures and structures of ligands bound to proteins, which have been discussed in section 1.5.2.1. For simplicity and in order to maximise the size of the data set, all complete data will be used in this analysis and ligands from structures with 2Å resolution or lower will be marked separately. Low resolution structures may thus be easily identified and incorporated into the conclusions with caution where necessary.

This chapter aims to investigate the extent of conformational variability and solvent accessibility of organic cofactors bound to proteins, in order to determine how it compares to the conformational variability of the same molecules in solution, whether the conformations are similar in homologous enzymes, how the results confirm, complement or contradict the study by Stockwell & Thornton (2006). In this chapter, it will be investigated if their conclusions hold true for a larger, more diverse set of ligands that are all catalytically active as cofactors. Please note that, in this work, the conformational variability is compared to that generated *in silico* to mimic as far as possible the true conformation in solution, while Stockwell & Thornton (2006) compared to a representative sample of all theoretically possible conformations. The impact of the results on the current knowledge of enzyme evolution and the validity of common docking-based methods in drug design will be discussed.

5.2 Conformational variability and solvent accessibility of cofactors

The catalytic power of some cofactors may depend on specific conformations in order to perform the required function. Here, it will be investigated whether all cofactors only adopt a certain (low) number of ideal conformations or whether they occupy most of the theoretically accessible low-energy conformational space. The aim is firstly to elucidate if and to what extent the observed cofactor conformations from the PDB crystal structures vary from the conformational space that each cofactor can theoretically adopt in solution, on all the available structural data from the PDB, as well as on a data set where representatives of related proteins are selected to reduce redundancy due to homology. Secondly, it shall be investigated whether the conformation of the cofactors depends on the homologous superfamily of enzymes they bind, *i.e.* whether the conformation of the cofactor is “conserved” within a homologous superfamily, identified by its CATH code. Thirdly, the relative solvent accessibility (see section 2.1.4 and equation 2.1) is analysed to reveal how deeply cofactors are buried, whether the level of solvent accessibility is similar for all atoms of the cofactors and whether general trends are observed in the data.

To assess the extent of conformational variability visually, all conformers of each cofactor are superimposed on a manually selected rigid portion of the molecule, where possible (see section 2.1.3). If there are several possibilities, the functional group (e.g. the flavin 3-ring system in FAD) was chosen as the basis for superposition.

The data set of protein-bound three-dimensional conformations has been gathered from the PDB. No equivalent experimental data were available for the conformations of the cofactor molecules in solution. The Cambridge Crystallographic Data Centre (CCDC, Allen, 2002) provides crystal structures of half a million small molecules. However, this resource usually only holds one crystal structure of each molecule, which thus cannot be used to obtain an overview of the conformational space that a molecule may adopt. Instead, a conformation generation program (OMEGA by OpenEye Software, OpenEye Scientific Software, 2004–2010) has been employed to generate a representative set of three-dimensional

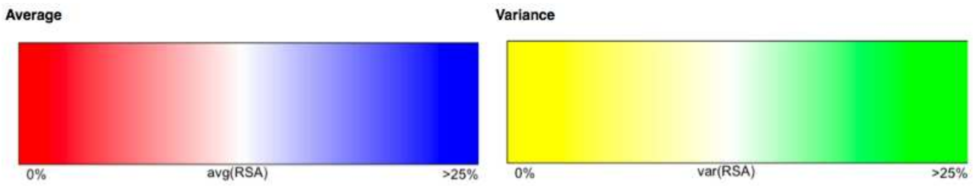


Figure 5.1: Legend for all the solvent accessibility figures in this chapter. The relative solvent accessibility for each atom is calculated following equation 2.1.

conformations in a force field that simulates a solution environment (Merck Molecular Force Field) to obtain low-energy conformers.

In order to assess the conformational variability mathematically rather than visually, histograms of the pairwise RMSD distance matrix of both data sets have been plotted in the same graph. Comparing the shapes of these distributions provides measures for the similarity of the conformational variability within and between the two data sets.

This analysis has been performed on the following cofactors (identified by their CoFactor PDB HET groups): ASC, BTN, COA, FAD, FMN, H4B, MGD, MTE, NAD, NAP, PLP, PMP, PQQ, SAM and TPP. Please note that the force field used by OMEGA does not support molecules that contain metal atoms. Running it without the metal atom or after replacing the metal atom by a supported atom type resulted in chemically unrealistic conformations. Thus all metal-containing cofactors (the porphyrin-like group plus MSS) are excluded from this analysis.

To assess whether the conformational variability of a ligand depends on the homologous superfamily of the enzyme to which it binds, the average pairwise RMSD between all[1] protein-bound structures of a cofactor was calculated, as well as the average for each homologous superfamily, identified by its CATH code.

As part of the CoFactor database pipeline (see figure 2.1 and section 2.1.4), the average and variance of the solvent accessibility of each ligand atom has been calculated. These results will be integrated in the discussion of the conformation results below. Please note that the legend that refers to all solvent accessibility plots is shown in figure 5.1.

[1] all those PDB entries for which the CATH code of the cofactor-binding domain is known

The results are visually more accessible in a three-dimensional molecule viewer than in printed two-dimensional figures. Thus, result web pages were created under http://www.ebi.ac.uk/thornton-srv/databases/jfischer/confvar using Jmol applets (Hanson, 2010) as well as RMSD distribution histograms. It is worth reiterating that, in the three-dimensional applets and screen shots thereof, the superposition was performed on aromatic or planar portions of the molecules where possible, in order to maximise the readability of the results. Yet, all the numeric calculations (such as the histograms, distribution plots as well as the averages and standard deviations of pairwise RMSD within homologous superfamilies) were performed on the superposition of the molecules on all their atoms in order to avoid a bias of the data towards the portion selected for superposition. The results are presented according to the groups of molecules determined in figure 4.11D.

5.2.1 Group 1: small cofactors

From group 1, the small cofactors, the data set comprises data for the HET codes ASC, BTN, PQQ, PLP, PMP and TPP. PMP was not part of the data set used in chapter 4, as the difference between PLP and PMP is only the NH_2 group that PLP transfers. As described in section 2.2.1.1, only one structure for each cofactor was selected there. Here, however, the data set is very small already. Therefore PMP has been included.

5.2.1.1 PQQ

PQQ is the simplest case, as this molecule is the most rigid with the lowest percentage of rotatable bonds (see table 4.1). In fact it is so rigid that OMEGA cannot produce two chemically realistic conformers that have an RMSD of greater than 0.5Å, as is shown in figure 5.2. Thus, the RMSD histogram cannot be plotted as there is only one generated conformer. Since there is such little conformational freedom in PQQ, only one conformation is adopted by all the conformers, in the protein as well as in the generated ones (see figures 5.2A and B).

The data set contains domains with three different CATH codes *i.e.* it contains structural information for the conformation of PQQ for three non-homologous su-

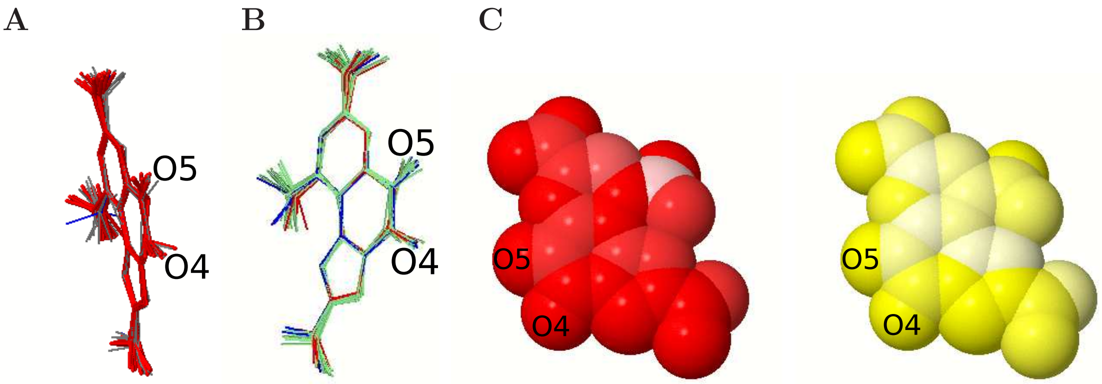


Figure 5.2: Conformations of PQQ. A: Protein-bound (red), protein-bound low resolution (grey) and computer-generated (blue). B: Protein bound structures coloured by CATH domain. C: Average and variance in solvent accessibility. The O4 and O5 atoms are catalytically active.

perfamilies of PQQ-binding domains. Since PQQ adopts only one conformation, the low-resolution structures agree with the high-resolution ones. The only conformational variability is observed for the catalytically active O4 and O5 atoms, which are not always in the ring plane (see figure 5.2A). Assuming that the ligands are modelled correctly, this non-planar conformation is most likely due to the catalytic atoms being in the fully reduced (sp_3 hybridised, both are alcohol groups) state as opposed to the fully oxidised (sp_2 hybridised, both are keto-groups) state, in which the atoms are expected to be located in the ring plane (see section 3.2.22).

The overall average RMSD distance for the 20 PQQ ligands shown in figure 5.2B is exceptionally low at 0.42±0.16Å. Three PQQ-binding CATH domains have been identified. Domain 1.20.910.10 and 2.120.10.30 have only three structures each and are thus too small to draw significant conclusions. The third CATH domain is 2.140.10.10, for which there are 14 examples in the PQQ data set, and has an average RMSD of 0.50±0.13Å. This is a very low value indicating that the conformation of PQQ within this superfamily is very restricted. However, there is very little difference to the overall average RMSD of PQQ, due to the low overall flexibility of the molecule.

The solvent accessibility analysis (see figure 5.2C) shows that the whole cofactor is extraordinarily buried, which is to be expected due to the largely hydrophobic character of the molecule (logP=0.55, see table 4.1). The O4 and O5 atoms with 0±0% solvent accessibility are tightly bound by the enzyme or a reactant in all structures.

It may be concluded that PQQ is a very rigid molecule and therefore generally does not vary a lot in the conformations it adopts. It is also deeply buried in the binding site.

5.2.1.2 PLP and PMP

For PLP and PMP, the data show that the conformational space of the protein-bound cofactors largely overlaps with the one adopted by the OMEGA-generated conformers (see figure 5.4A). However, figure 5.4C illustrates that, in the protein-bound conformation ensemble, the phosphate group in PLP (and PMP, not shown, please refer to the result web pages) is always orientated in a *trans*-like conformation on the O4P-C5A bond, while OMEGA generates both *cis*-like and *trans*-like conformations at this bond. In the *trans*-like conformation, the O4P atom points away from the ring system, whereas in the other two low-energy dihedral angles for this bond, the O4P atom is located closer to the ring system (*cis*-like, less extended conformation). This may be seen as evidence supporting the hypothesis that cofactors are more extended when bound to a protein than in solution.

Browsing all the known mechanisms for PLP in the MACiE database, one can observe that the PLP-ring is held in place by a hydrophobic residue in the active site of most PLP-enzymes. This is mechanistically important as it prevents the cofactor from rotating around the C5-C5A bond.

The PMP and PLP histograms (see figure 5.3) confirm these findings: while the protein-bound data sets exhibit a rather smooth, unimodal distribution, the generated ones show a right shifted, flatter, unimodal distribution, which confirms that the conformers are more evenly spaced and that a larger conformational space is adopted.

The CATH filtered data for PLP appears to be representative (see figures 5.4B and 5.3B) since the histogram as well as the data displayed in the Jmol applet

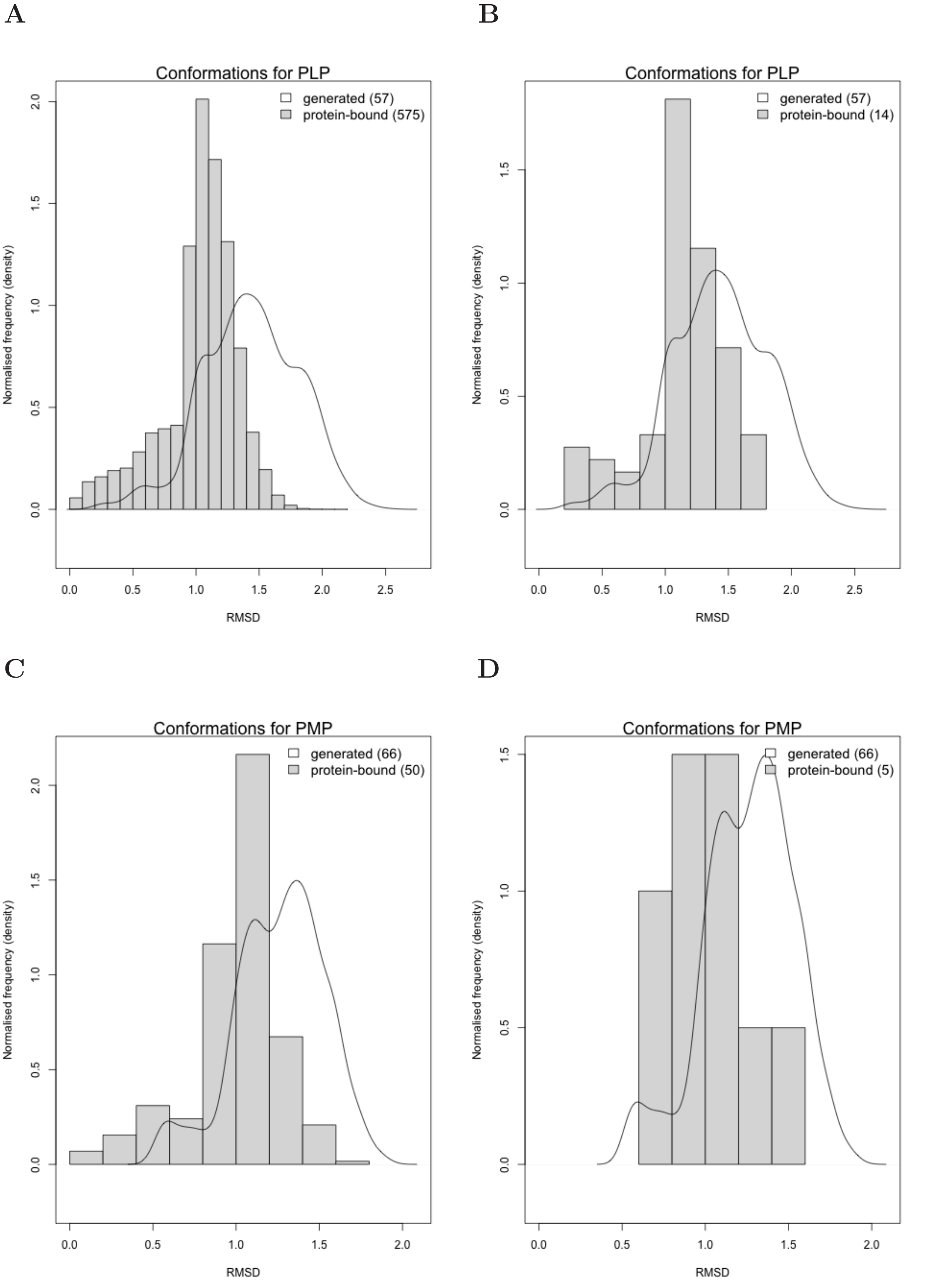


Figure 5.3: Density histograms and distributions of PLP (top) and PMP (bottom), based on the structures generated by OMEGA superimposed on the full data set (left) and the homology-filtered data set (right).

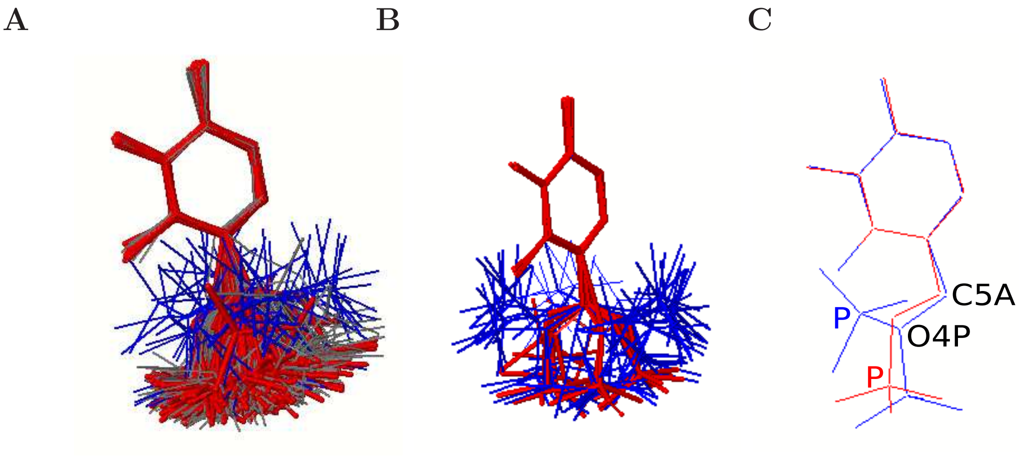


Figure 5.4: Conformations of PLP, protein-bound (red ≤2Å resolution, grey >2Å), and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. C: *cis*- and *trans*-like conformations of the C5A-O4P bond.

largely agree with the unfiltered data. The same is true for PMP (not shown, see web site), although less obvious as the data set after filtering is very small. In the case of PLP, this indicates that the cofactor is indeed more extended in the protein structures than in the generated ones.

The homology analysis shows that the conformation of this cofactor does not depend on the domain family. Figure 5.5 shows all PLP ligands from the PDB coloured by CATH code (A), followed by the five most populated CATH families in order of population size (B-F: 114, 108, 104, 54 and 19, respectively).

Figure 5.5B-E show PLP superpositions from CATH families with at least 54 examples. Their average RMSDs are 1.14±0.27Å, 1.17±0.27Å, 1.90±0.78Å and 1.19±0.78Å, respectively. The next smaller CATH family ligand set has only 19 examples (figure 5.5F) and only then there is a noticeable difference, in the figure as well as in the numbers. With this last family's average RMSD at 0.60±0.34Å, this is the first sign of conformational restriction compared to the full data set (1.42±0.53Å, sample size 451 structures, figure 5.5A), although this might be due to lack of data (only 19 examples).

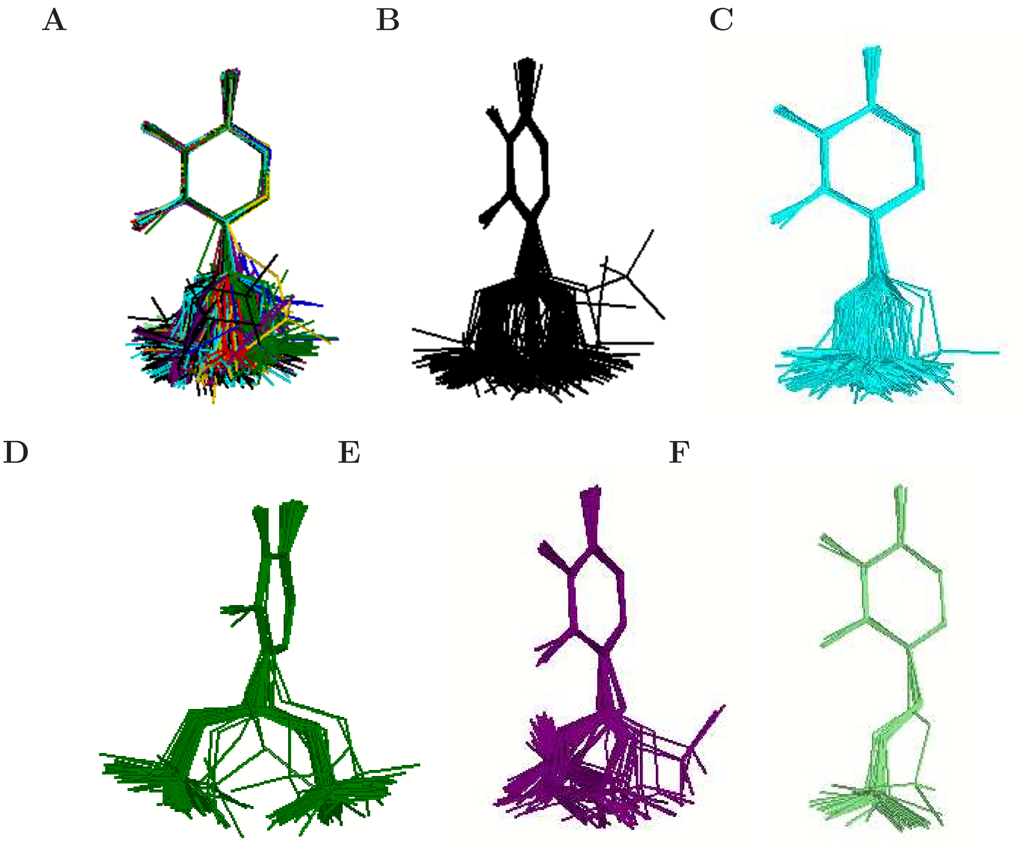


Figure 5.5: Conformations of protein-bound PLP structures. A: coloured by CATH code, B: the 114 examples of CATH family 3.40.640.10, C: the 108 examples of CATH family 3.90.1150.10, D: the 104 examples of CATH family 3.40.50.2000, E: the 54 examples of CATH family 3.40.50.1100, and F: the 19 examples of CATH family 3.20.10.10.

The solvent accessibility visualisation shown in figure 5.6 indicates that PLP is very tightly bound by its host enzymes. In combination with the results from the previous paragraph, this may be interpreted in two ways. Either, the binding site varies slightly in enzymes of the same CATH family: indicating that the cofactor binds tightly enough to explain the low average and variance in solvent accessibility, but to still allow for the intra-family conformational variation seen in figure 5.5; or the other molecules that take part in the enzyme reaction might occupy the binding site in different ways so that PLP's low solvent accessibility but high partial conformational variability may be explained. Possibly, both of these explanations contribute to the observed results.

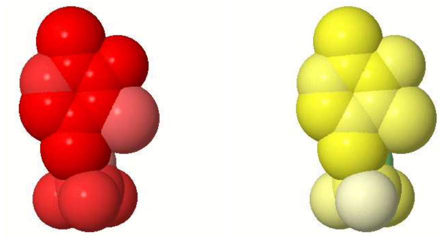


Figure 5.6: Average (left) and variance (right) in PLP solvent accessibility in the PDB.

Summarising all observations about PLP, it cannot be concluded for this cofactor that its conformation depends on the homologous superfamily to which it is bound. The variability is however mostly due to a flexible and small phosphate group, which is in most cases not involved in catalysis.

5.2.1.3 TPP

The analysis of thiamine diphosphate (TPP) in figure 5.7 reveals that the protein-bound conformations are more stretched out than the generated ones, for both the full and the CATH filtered data set (see figure 5.7A and B).

The superposition was performed on the three linker atoms between the two rings in the molecule. The two bonds between these three atoms are rotatable single bonds. Figure 5.7C reveals that the orientation of the two rings towards one another is similar for all of the molecules in the data set: the N4’ atom of the aminopyrimidine ring is usually oriented on the opposite side of the molecule to the CM4 atom of the thiazolium ring. The latter ring determines on which side of the molecule the side chain is located. Two conformational clusters arise from this flexibility, depending on where the side chain is located. The effect of these spatial clusters is seen in the distribution plots (see figure 5.8A and B), where the two peaks in the histograms correspond to the two spatial clusters. The generated structures show a more evenly spaced range of conformations, resulting in one peak in the distribution plots.

Figure 5.7E shows only the protein-bound TPP data set, where the low-resolution structures are shown in grey. Once these data are removed (see figure

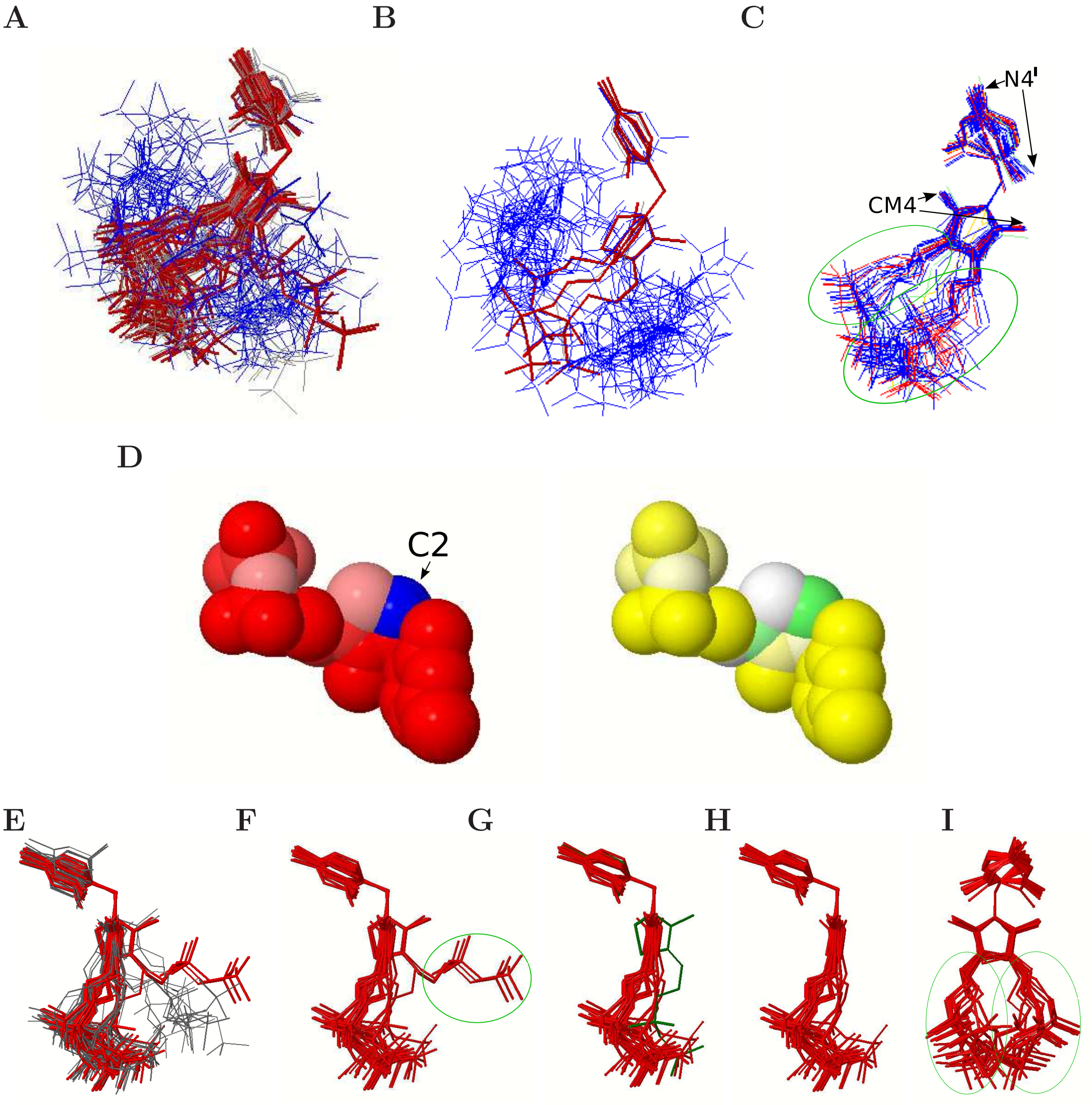


Figure 5.7: Conformations of TPP, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. C: Conformations coloured by CATH code, and D: Average (left) and variance (right) in TPP solvent accessibility. E: Only the protein-binding structures. F: Only the high resolution protein-binding structures. G: Like F but without the PDB codes 3e78, 3e79 and 3eki published by Sippel *et al.* (2008). TPP from PDB code 3c9u highlighted in green. H: Like G but without the PDB code 3c9u from the thiamine-monophosphate kinase. I: like H but rotated by 90deg to show the two structural clusters (green circles).

5.7F), three obvious outliers can be identified (green circle in figure 5.7F). All three of these crystal structures have been published with the same study (Sippel *et al.*, 2008). The authors write in the abstract that "An unanticipated ligand bound in the centre of the molecule at the base of the cleft has been modelled as thiamine pyrophosphate or vitamin B(1)." This implies that the ligand may not actually be TPP and should thus be excluded from the analysis. When removing these three structures, figure 5.7G is obtained. One of the remaining structures (green in figure 5.7G) has a different conformation to the other ones. Following the PDB code (3c9u) reveals that this protein is a high-resolution structure of the last enzyme in TPP biosynthesis (thiamine-monophosphate kinase, no E.C. number assigned). This may explain the difference in conformation as the cofactor is not required to participate actively in catalysis. The remaining data set is homogeneous and is shown in figures 5.7H and 5.7I (I rotated by 90 degrees to show the two clusters mentioned above).

The solvent accessibility analysis (see figure 5.7D) shows that TPP is generally very tightly bound with the exception of the catalytic C2 atom, which needs to be accessible to the substrates.

5.2.1.4 ASC

For ascorbic acid (ASC), the data set does not contain a lot of information since there are only nine ASC-containing PDB entries deposited that were obtained using X-ray crystallography. At least most of the data is of high quality, as only two of these structures have been solved at 2Å resolution or worse.

As ASC is a small and ring-containing molecule, the rotational freedom is limited by a small number of rotatable bonds. The C4 atom (see figure 5.9A) is the stereo centre that distinguished L-ascorbic acid from D-ascorbic acid and therefore adopts the same conformation in all L-ascorbic acid molecules. In the generated data, the C6 atom occupies all three energetically favourable positions, whereas this atom's position is less variable in the data from the PDB. In the CATH filtered data there is no low resolution data. Thus, the superposition analysis (see figures 5.9A and B) as well as the distribution analysis (see figure 5.8C and D) indicate that the conformation of ASC in the protein-bound data is

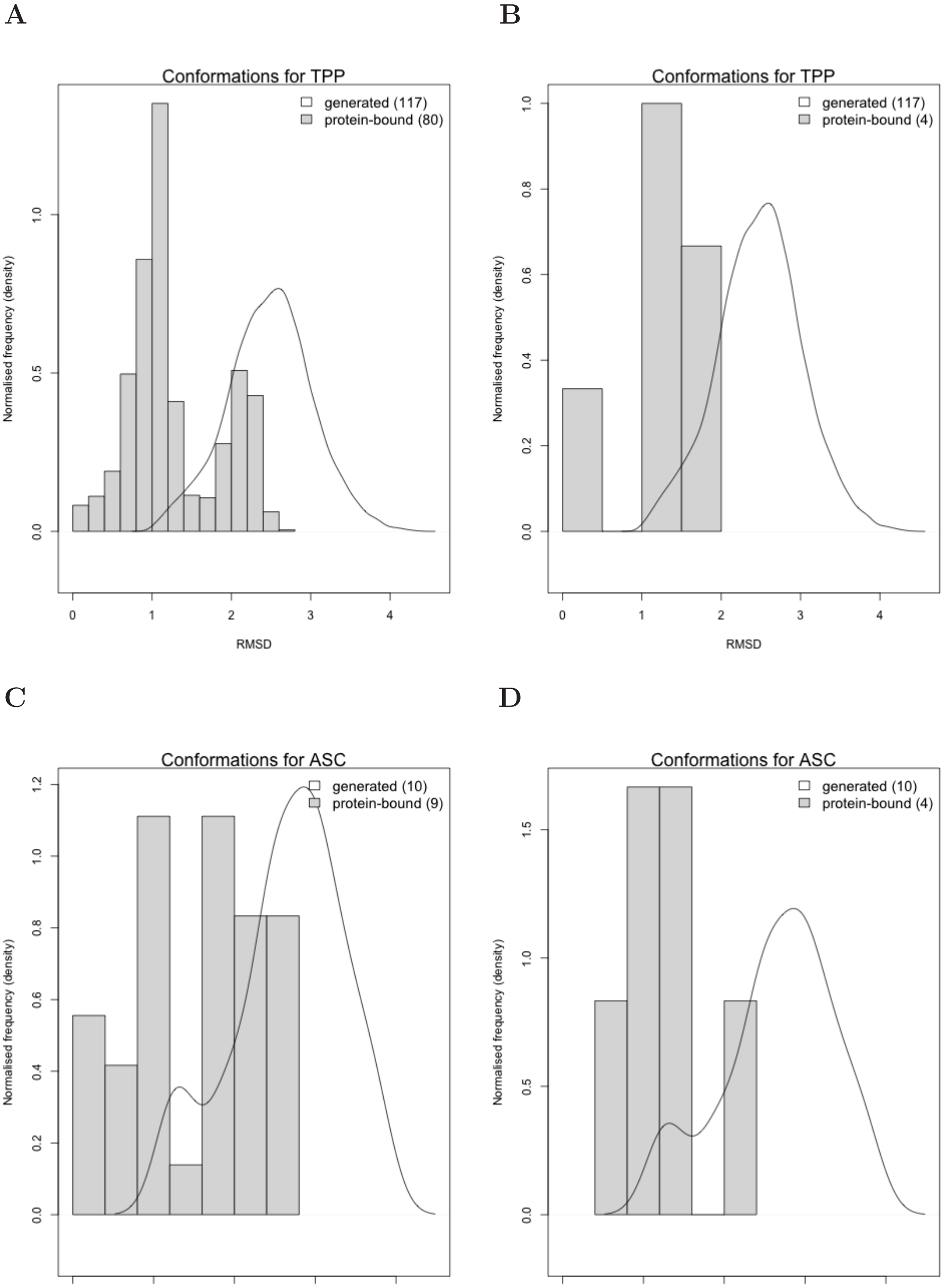


Figure 5.8: Density histograms and distributions of TPP (top) and ASC (bottom), based on the structures generated by OMEGA superimposed on the full data set (left) and the homology-filtered data set (right).

slightly more restricted then in the generated data. Please note, however, that this might be due to the very small sample size of the data set, although it does cover three different homologous superfamilies.

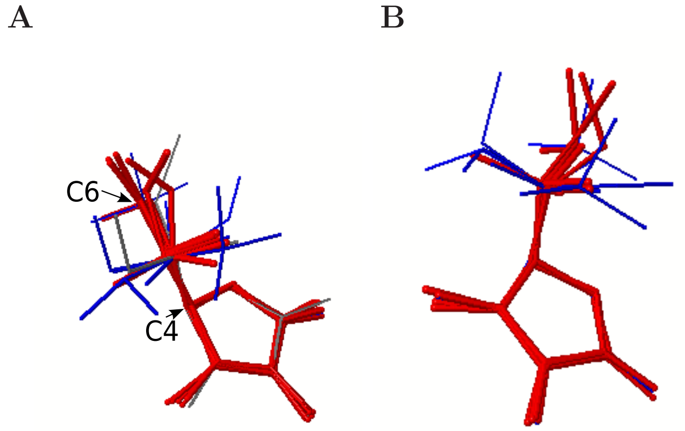


Figure 5.9: Conformations of ASC, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures.

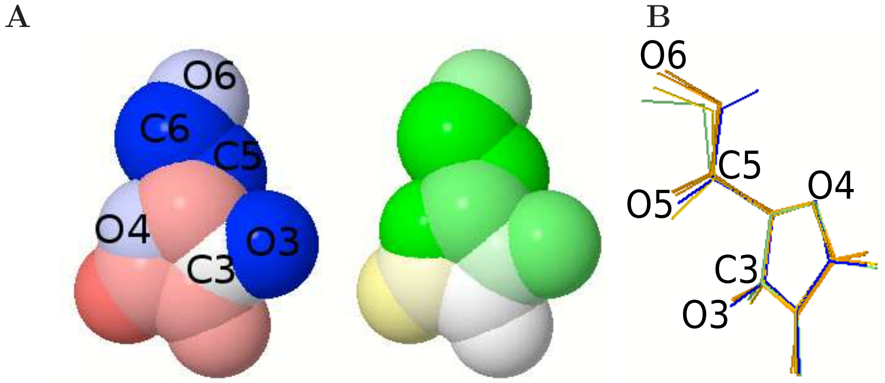


Figure 5.10: A: Average (left) and variance (right) in ASC solvent accessibility in the PDB. B: Conformations of protein-bound ASC structures coloured by CATH code.

The homology analysis shows that there are 4 different CATH families represented in the ASC ligand data set (see figure 5.10B). Three of those are only represented with one structure each and the homologous superfamily identified by CATH code 3.20.20.80 has three representative structures with an average RMSD of 0.08±0.002Å, as compared to the 0.47±0.33Å between all 6 structures that have a CATH id assigned. Unfortunately this sample size is far too small to draw any significant conclusions.

Figure 5.10A shows the visualisation of the average and variance in solvent accessibility for the ASC structures. Compared to most of the other cofactors, ASC exhibits a higher solvent accessibility for some atoms: C3, O4 and O6 are at least 10% accessible and O3, C5, C6 and the hidden O5 are at least 25% exposed to the solvent. This is noted as an interesting exception, but the same restrictions imposed by the small sample size prohibit strong conclusions.

5.2.1.5 BTN

The last member of group 1 cofactors in this data set is biotin (BTN). The superposition analysis (see figure 5.11A) indicates that, for this molecule, only a small subset of the generated conformers are actually adopted when it is bound to a protein. The double ring system adopts an angled position at the C4-C5 bond (see figure 5.11C): the nitrogen-containing ring shows a planar conformation in the PDB models whereas the sulphur-containing ring is always modelled to adopt a boat conformation with respect to the other ring plane.

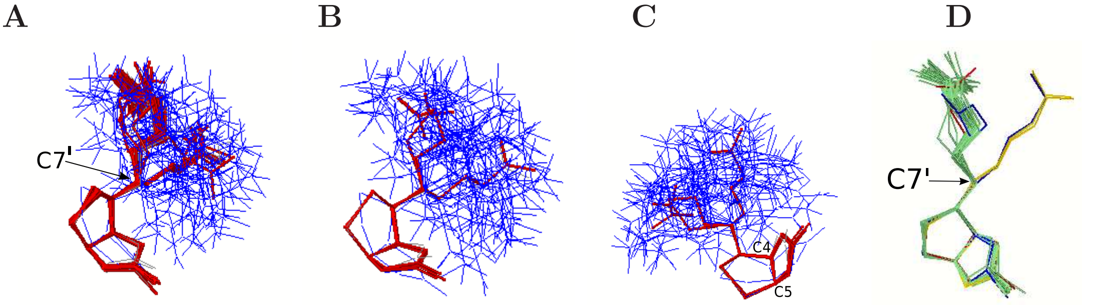


Figure 5.11: Conformations of BTN, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. C: Boat and chair conformations. D: Protein-bound structures coloured by CATH code.

In contrast, some of the generated conformers adopt a chair conformation. The molecules were superimposed on both rings and C7. This results in the effect seen in figure 5.11A-C: in one half of the generated structures the ring adopts the same boat conformation as in the PDB structures (hidden beneath the PDB structures), whereas the ring is in a chair conformation in the other half. Although the side chain has four rotatable C-C bonds, all protein-bound molecules are extended whereas the generated ones also adopt more compact conformations. This result is further confirmed by the left-shift of the protein-bound molecules compared to the generated ones in the distribution plots (see figure 5.13A and B). The CATH filtering leaves four structures, all of which are extended.

The homology analysis shows that, of the four superfamilies, only one is populated by several (37) structures (2.40.128.30, mint in figure 5.11D). This family exhibits a slightly restricted conformational variability of 1.34±0.03Å pairwise RMSD, compared to all CATH-labelled structures at an RMSD of 1.36±0.12Å. Although there are only two structures for CATH code 3.30.930.10, it is worth noting that these two structures adopt opposite conformations at C7', unlike the previously discussed family.

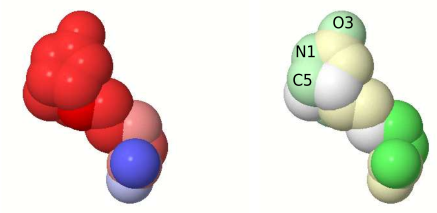


Figure 5.12: Average (left) and variance (right) in BTN solvent accessibility in the PDB.

Figure 5.12 shows the average and variance in solvent accessibility of protein-bound biotin structures. The side chain is less tightly bound than the ring system, but overall, the solvent accessibility is high. It is worth noting that the solvent accessibility is more variable at the N1 and its neighbouring O3 and C5 atoms, possibly due to the substrate binding there.

5.2.1.6 General observations for the small cofactors

In general, protein-bound cofactors from group 1 exhibit a tendency to adopt more extended conformations compared to the computer-generated low-energy distribution. This extended conformation depends to varying degrees on the global flexibility of the molecule. Yet, the difference between the generated and the protein-bound structures is small, most likely due to the small size of the molecules. Most of the cofactors in this group are deeply buried in the binding protein domain(s) and the catalytic atoms are on average more exposed to the solvent then the other atoms. A conformational dependency on the homologous superfamilies is not obvious in most cases, although a slightly restricted conformational variability is observed for biotin.

5.2.2 Group 2: mono-nucleotides

Group 2 cofactors are those that may be roughly described as mono-nucleotides. From this group, conformational data has been analysed for the PDB HET codes COA, FMN, H4B, SAM and MTE.

5.2.2.1 SAM

Starting off with SAM, the superposition applet shows that there are two very common clusters of conformations (see figure 5.14C), which are distinguished mainly through the torsion angle between the adenine plane and the ribose plane, rotating around the N-glycosidic N9-C1' bond. Those two conformational clusters both consist of extended conformers, where the methionine side chain points away from the adenosine residue. Contrary to the trend observed for most cofactors, some models of SAM deposited in the PDB show the methionine "folded back" towards the adenine portion (only towards C8, never towards N3, see figure 5.14A). However, the two-cluster trend is strong enough to be documented in the corresponding histogram (see figure 5.13C), where once again two peaks can be identified, one for the RMSDs between the conformers from the same cluster and one measured between both clusters. Although the first peak at 0.75Å is caused by all those structures that are very similar to each other, please note

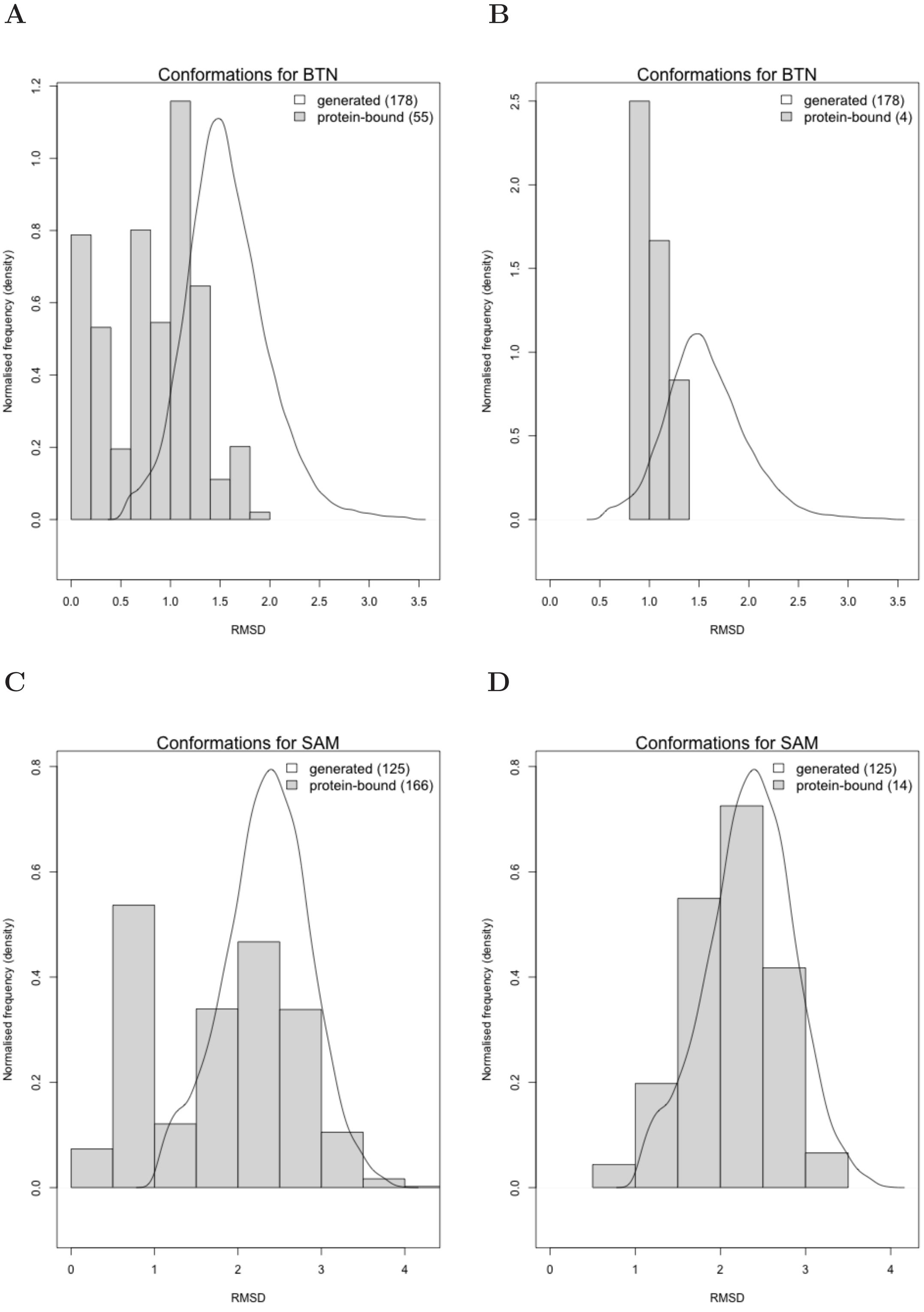


Figure 5.13: Density histograms and distributions of BTN (top) and SAM (bottom), based on the structures generated by OMEGA superimposed on the full data set (left) and the homology-filtered data set (right).

that figure 5.14C shows all the SAM structures coloured by CATH code. This reveals that all of the superfamilies with at least 4 structures occupy both of the conformational clusters. The double peak in the histogram disappears in the CATH-filtered histogram (see figure 5.13D), as it comprises only four superfamilies and the best-resolution representatives are more evenly spaced.

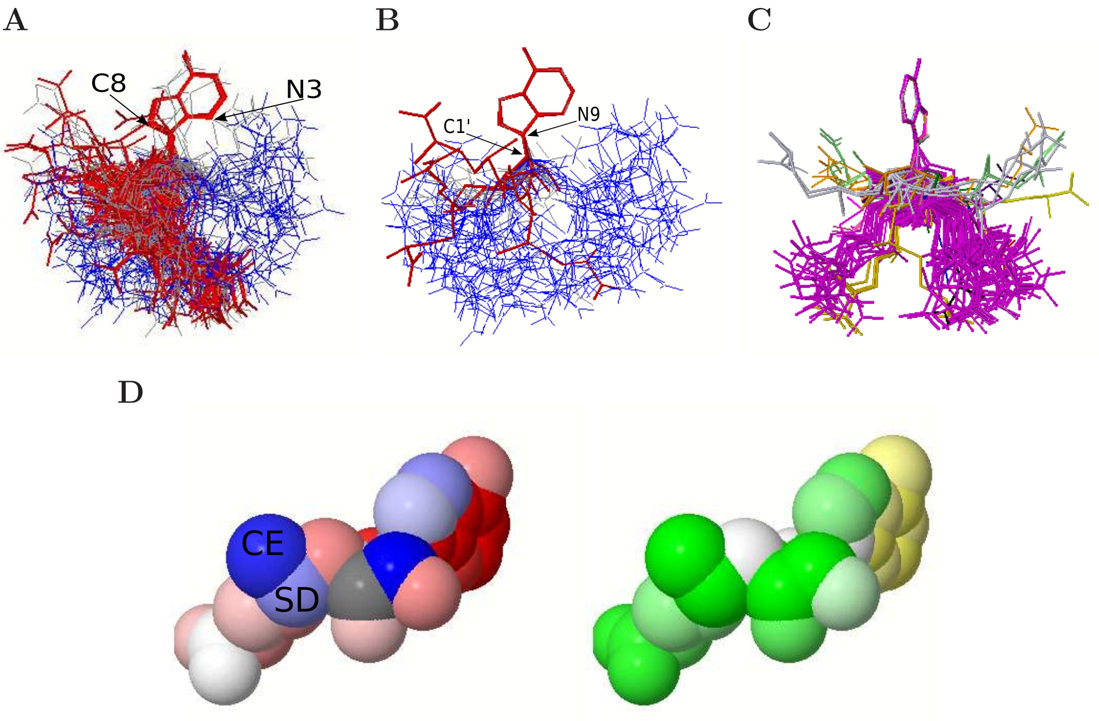


Figure 5.14: Conformations of SAM, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. Protein-bound structures that have CATH codes assigned coloured by CATH code (C). Average and variance in relative solvent accessibility of protein-bound SAM structures (D)

The conformers generated by OMEGA naturally do not show this trend because they are filtered for mutual RMSD and also because the folding back to N3 does occur for the generated, but not for the protein-bound structures. This fact is mirrored in the distribution line for the generated structures (see figure 5.13D). Although a large part of the conformational space that SAM occupies in solution (as predicted by OMEGA) is observed in the protein-bound models, it is worth noting that, in this data set, the two main conformational clusters are much more densely populated than the remainder of the conformational space.

When the data are filtered by CATH codes, one can observe that most of the conformational space seen in the full data set is still occupied here (see figure 5.14B), although the histogram (see figure 5.13D) only shows one peak due to the reduced data set. However, while the full data set is much more densely populated in the extended conformations, the unique homologous superfamily data set shows that this is due to a bias towards certain CATH families: 3.40.50.150 (magenta), 3.30.300.10 (mint), 3.30.46.10 (yellow) and 3.40.1280.10 (orange, all see figure 5.14C). Interestingly, all these families occur in both of the aforementioned structural clusters. The first of the families (3.40.50.150, magenta) comprises 52 structures and shows an average pairwise RMSD of 2.63±1.26Å, compared to the 6.08±1.78Å of the full data set (80 structures). This suggests that the conformation is dependant on the homologous cofactor in the case of SAM.

The solvent accessibility analysis reveals that the catalytically active atoms (SD atom, and in group transfer reactions also CE atom, see figure 5.14D) are more exposed than most of the other atoms in the cofactor. The adenine ring is mostly buried.

5.2.2.2 COA

Coenzyme A (COA) has a high percentage (60%) of rotatable bonds and is composed of pantothenic acid as well as other parts. As mentioned in chapters 3 and 4, pantothenic acid's function includes mobility, *i.e.* to pass a substrate or intermediate between active sites. This function requires the flexibility of pantothenic acid, and it follows that COA's conformations may be more variable than the average cofactor's, which is confirmed in figure 5.15A. OMEGA has generated an ensemble of 13 conformers. Because COA is so highly flexible, this low number is most likely due to the energy cut-off.

The homology analysis of COA (see figures 5.15C and D) shows that clustering by homologous superfamily only occurs partially, and mainly in families with a small sample size. From the 200 COA structures in the PDB, of which 122 have been assigned a CATH code, the three most populated families comprise 25, 23

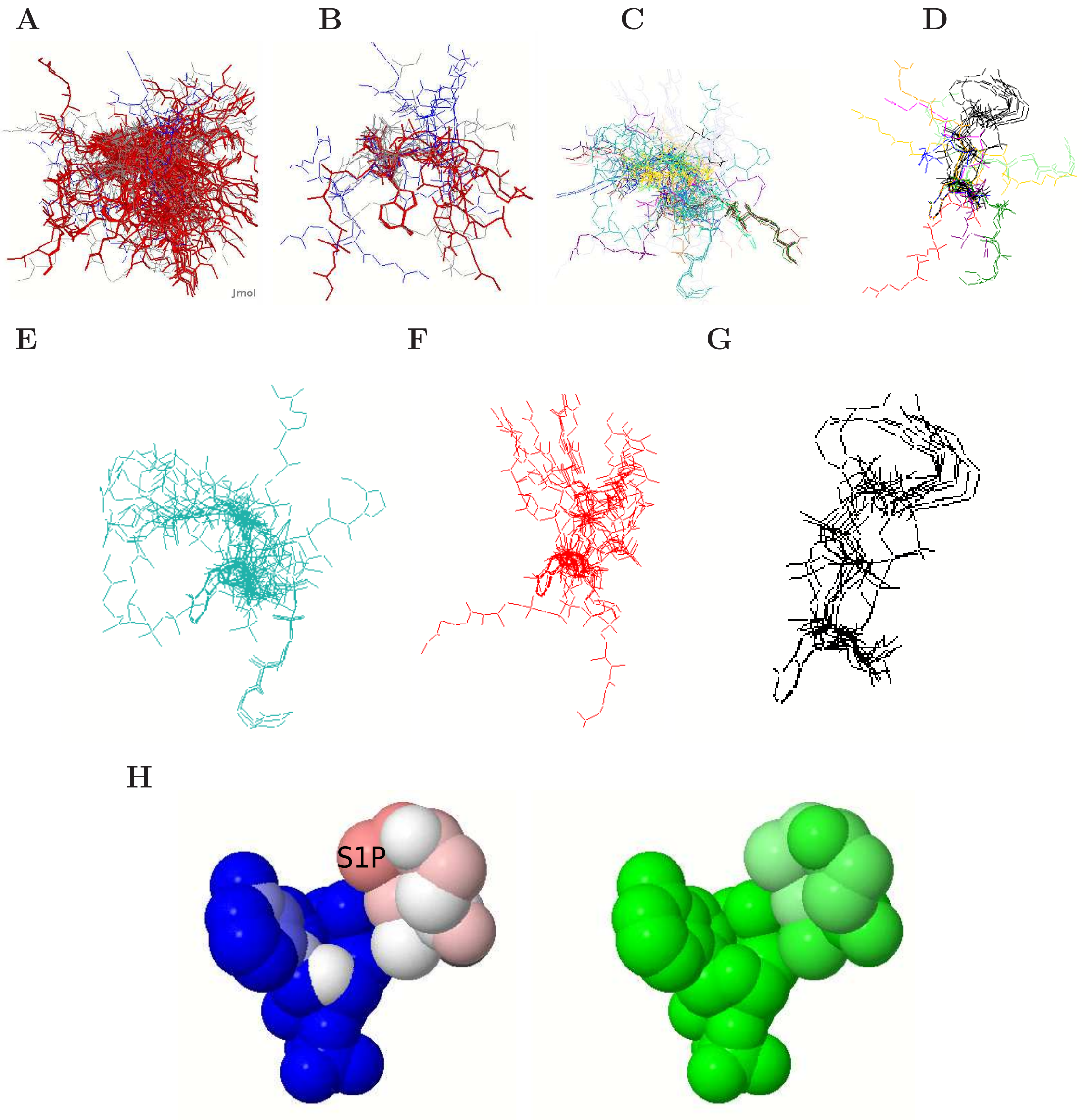


Figure 5.15: Conformations of COA, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the computer-generated structures. Protein-bound structures that have CATH codes assigned coloured by CATH code (C,D). COA structures bound in CATH families 3.40.630.30 (petrol), 3.40.47.10 (red) and 3.40.50.10540 (black). Average and variance in relative solvent accessibility of protein-bound COA structures (H)

and 11 structures, respectively, and are shown in figures 5.15E, F and G. These figures show that, although the structures in none of the three families occupy the whole of COA's conformational space, their conformations show little spatial restriction, at least in the case of figures 5.15E and F.

The above observations are consistent with the results from the solvent accessibility analysis (see figure 5.15H, where the catalytic sulfur atom S1P has been marked). The average solvent accessibility for this flexible cofactor is high (blue atoms in figure 5.15H: $\geq$25%). The variance in solvent accessibility is large for nearly all atoms in the cofactor (green atoms in figure 5.15H: $\geq$25%); only the catalytic S1P atom and its neighbours are slightly below that threshold. All together it appears that COA's conformation is less dependant on being bound to a particular superfamily than this is the case for other cofactors.

Again, the distribution plot shows a smooth curve with a single peak for the pairwise RMSDs between the generated structures. Based on the one-peak shape of both the histogram and the smoothed line in figure 5.16A, it can be seen that there is no major clustering into preferred conformations in the full set. The same is observed for the CATH-filtered data (see figures 5.15B and 5.16B). A left-shift of the PDB-based histogram compared to the generated distribution line can be identified in figures 5.16A and B. Hence the higher average RMSD among the generated structures (right shift compared to the histograms) implies that the conformational space is sampled more sparsely in the generated conformers than in the protein-bound models. The shoulder of the distribution in figures 5.16A and B is probably a result of the energy cut-off, rather than a real conformational restriction. As mentioned above, OMEGA has only generated 13 structures for COA. It may be concluded that COA is a highly flexible molecule in both data sets, which conforms with prior knowledge. No clear conformational clusters are observed.

5.2.2.3 FMN

The analysis of FMN structures indicates that this cofactor confirms the findings published by Stockwell & Thornton (2006). Figures 5.17A and B show that, in

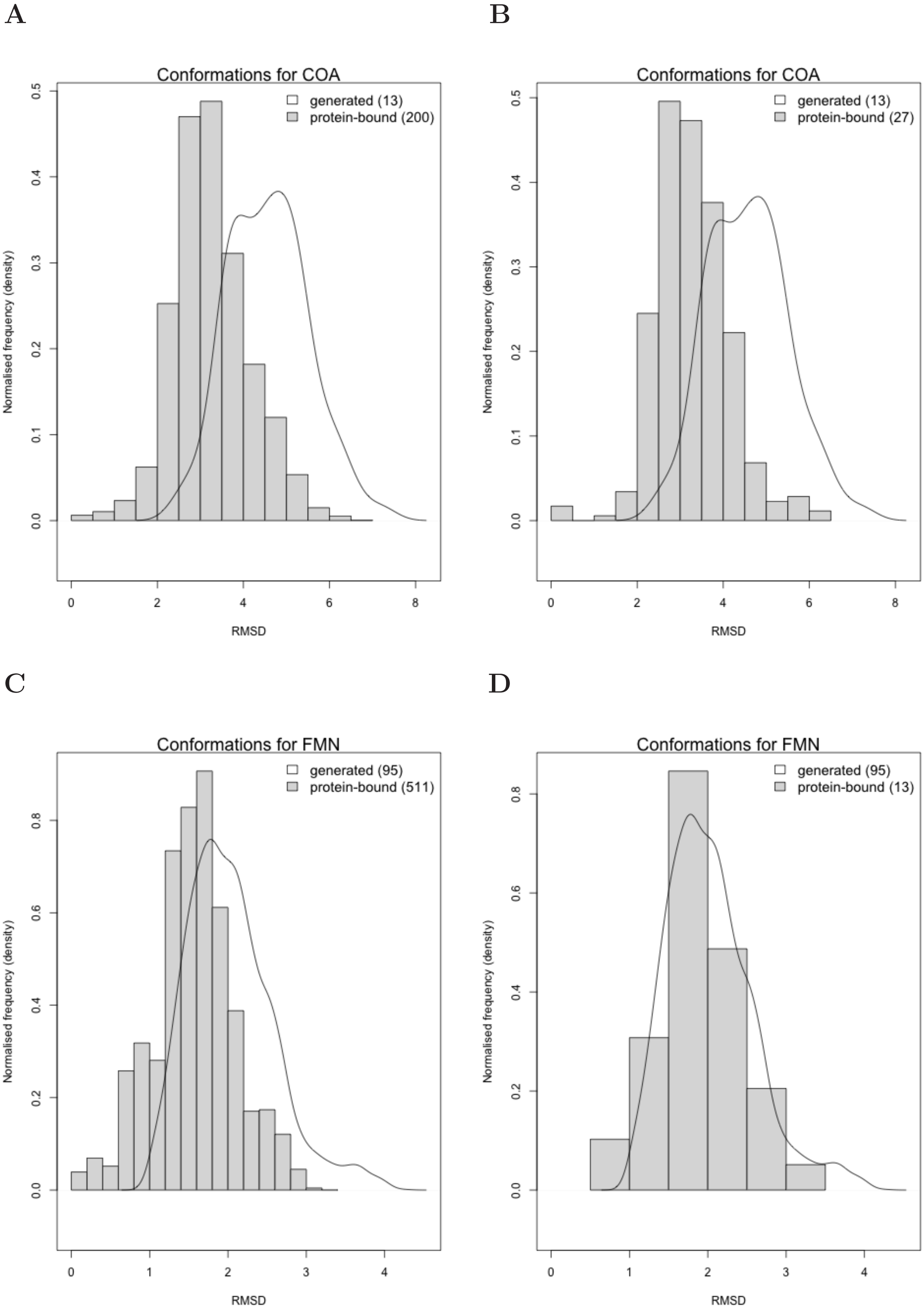


Figure 5.16: Density histograms and distributions of COA (top) and FMN (bottom), based on the structures generated by OMEGA superimposed on the full data set (left) and the homology-filtered data set (right).

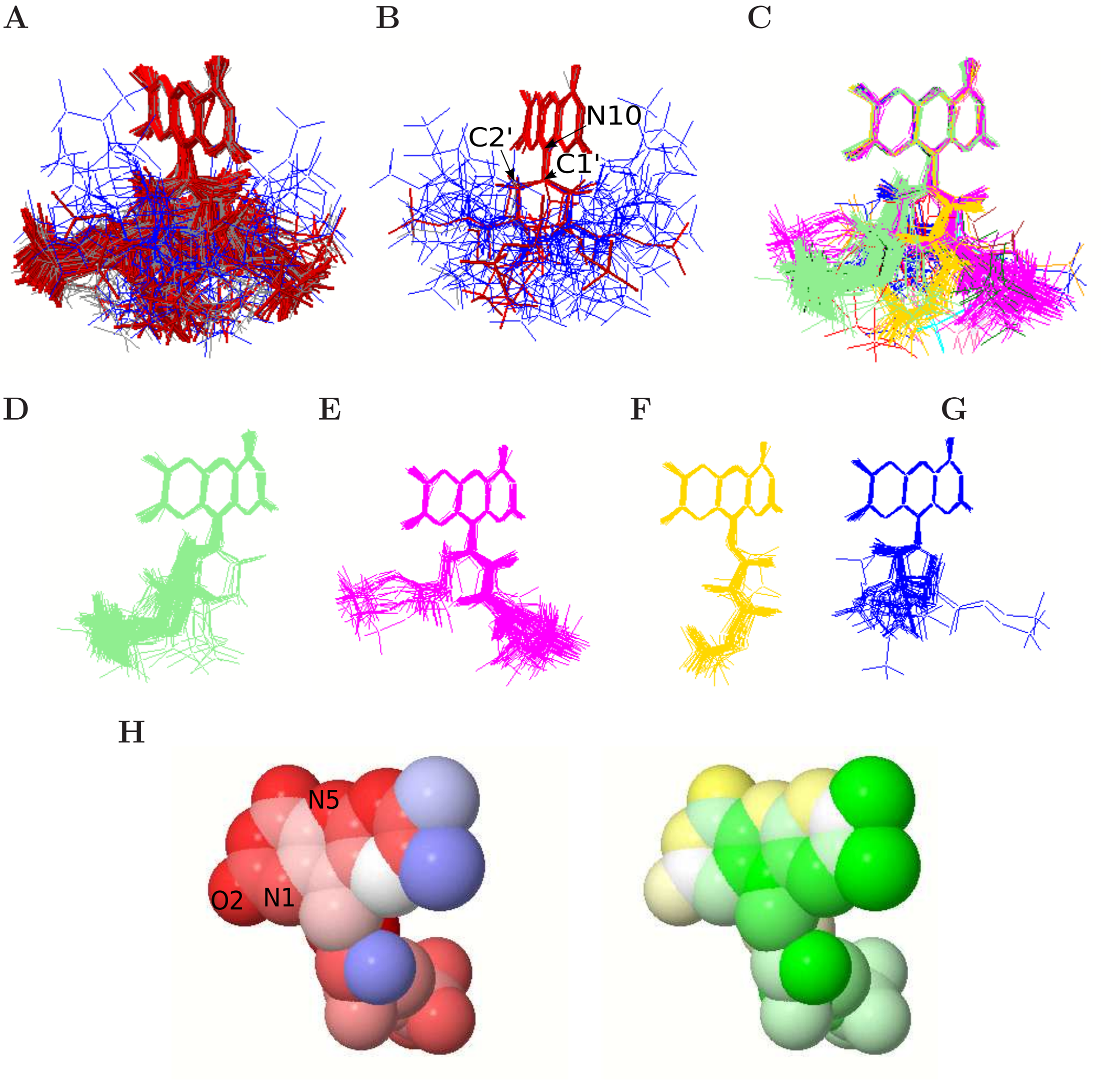


Figure 5.17: Conformations of FMN, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. Protein-bound structures that have CATH codes assigned coloured by CATH code (C). FMN structures bound in CATH families 3.20.20.70 (D, mint), 3.40.50.360 (E, magenta), 3.40.109.10 (F, yellow) and 2.30.110.10 (G, blue). Average and variance in relative solvent accessibility of protein-bound FMN structures (H).

the full data set as well as the CATH filtered one, the cofactor is generally more extended then in the OMEGA-generated structures.

The FMN distribution plot shows a smooth distribution with one peak for the generated structures (figure 5.16C) and a relatively smooth but slightly left-shifted histogram for the protein-bound ones (figure 5.16D). This shift is once again evidence for the increased amount of being compact, or "folded back" onto the flavin ring system in the generated structures. The filtering by homology drastically reduces the data set. The filtered protein-bound conformations produce a more symmetric and less left-shifted histogram which indicates a possible bias that may result from a too small data set after CATH filtering (11 structures).

Of the 341 FMN structures with 13 different CATH codes, figures 5.15D-G show the four most populated CATH families. While the structures in the full FMN data set have an average pairwise RMSD of 3.12±1.56Å, the four families have clearly lower averages, which indicates that the conformation of FMN is at least partially conserved within the homologous superfamilies. The family identified by CATH code 3.20.20.70 (figure 5.15D, mint) has an average pairwise RMSD of 1.10±0.50Å at a sample size of 139 structures. While most of the conformers are extremely similar to each other, some conformers adopt the opposite conformation at N10-C1'-C2', such that the C2' atom is in front of or behind the flavin plane. Superfamily 3.40.50.360 (figure 5.15E, magenta) has an average pairwise RMSD of 1.67±1.76Å at a sample size of 99 structures and clearly adopts both conformations at N10-C1'-C2'. 3.40.109.10 (figure 5.15F, yellow) has an average pairwise RMSD of 0.89±0.20Å at a sample size of 33 structures. The FMN conformation in this superfamily appears to be very restricted, and only one of the two possible conformations is adopted. Lastly 2.30.110.10 (figure 5.15G, blue) has an average pairwise RMSD of 1.99±0.87Å at a sample size of 32 structures. These structures adopt a more versatile set of conformations, not only at N10-C1'-C2' but also at other points in the ribophosphate moiety (see three-dimensional representation on web pages). This causes the high average RMSD value measured.

The relative solvent accessibility of FMN also follows the general trend. The cofactor is tightly bound with a relative solvent accessibility of $\leq 25\%$ for all and $\leq 12\%$ for most atoms, including the catalytic N1, O2 and N5 atoms, but excluding three others (see figure 5.15H). Overall, this indicates that FMN is very tightly bound in the active site, either by the protein or by the substrates.

5.2.2.4 MTE

In the case of MTE, again, the side chain conformation is slightly more restricted in the protein-bound conformers than in the generated ones. However, there are only 26 protein-bound MTE structures (see figure 5.18A) with 4 unique CATH codes (see figure 5.18B and C), so the restricted distribution may be caused by the small sample size.

The two peaks in the histogram in figure 5.19A result from the two clusters of side chain conformations shown in figure 5.18D. The two clusters arise from rotation about the rotatable C3'-C4' bond. The third low-energy conformation resulting from an sp_3-hybridised C-atom is not observed, which may be a sample size problem. The C4'-O4' bond is also rotatable and separates the grey and blue structures from the black ones. After filtering by CATH code, this histogram shows only one peak, corresponding to the cluster marked in figure 5.18E.

Figure 5.18C shows the protein-bound MTE structures coloured by CATH code. Visual inspection indicates that cofactor conformation depends on the homologous superfamily here, as the structures from the same CATH codes cluster together. However, the data set is too small to rely on the average RMSD calculations.

The solvent accessibility analysis (see figure 5.18F) reveals that MTE is extraordinarily buried with most atoms being less than 5% accessible to the solvent.

5.2.2.5 H4B

Lastly, biopterin (H4B) displays the same trend observed in the other mononucleotide cofactors. Again, the side chain conformation is more restricted in the protein-bound structures than in the generated ones (see figures 5.20A and B). In the fully reduced form of the cofactor, the C6 atom may be located on either side of the ring plane and determines in which of the two clusters the side chain is located (see figure 5.20C). These conformational clusters cause the two peaks in the histogram plot in figure 5.19C. The CATH filtered version (see figure 5.19D)

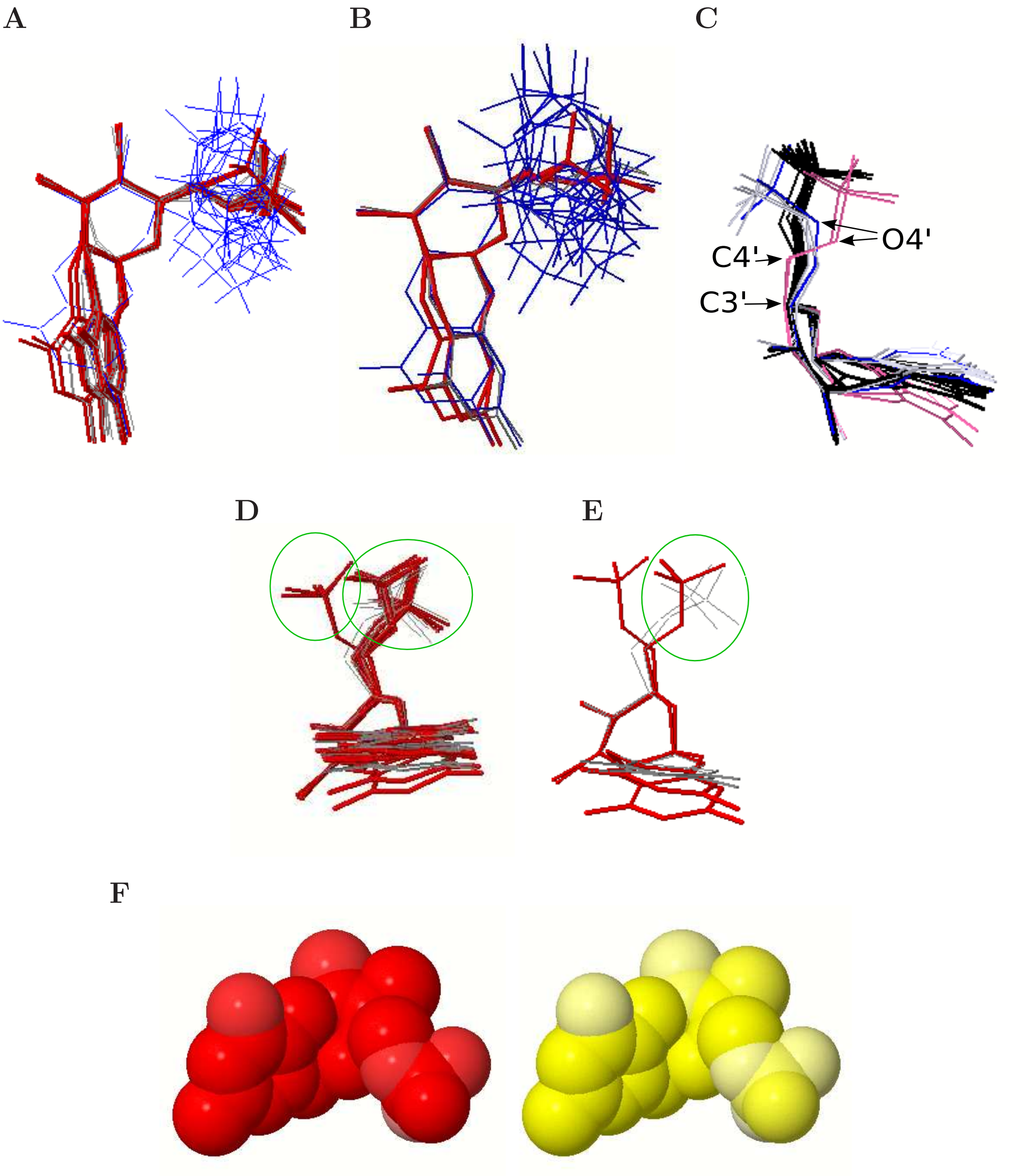


Figure 5.18: Conformations of MTE, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. Protein-bound structures coloured by CATH code (C). All protein-bound (D) and CATH-filtered (E) MTE structures showing the clusters (green circle) of the side chain conformers. Average (left) and variance (right) in relative solvent accessibility of protein-bound MTE structures (F).

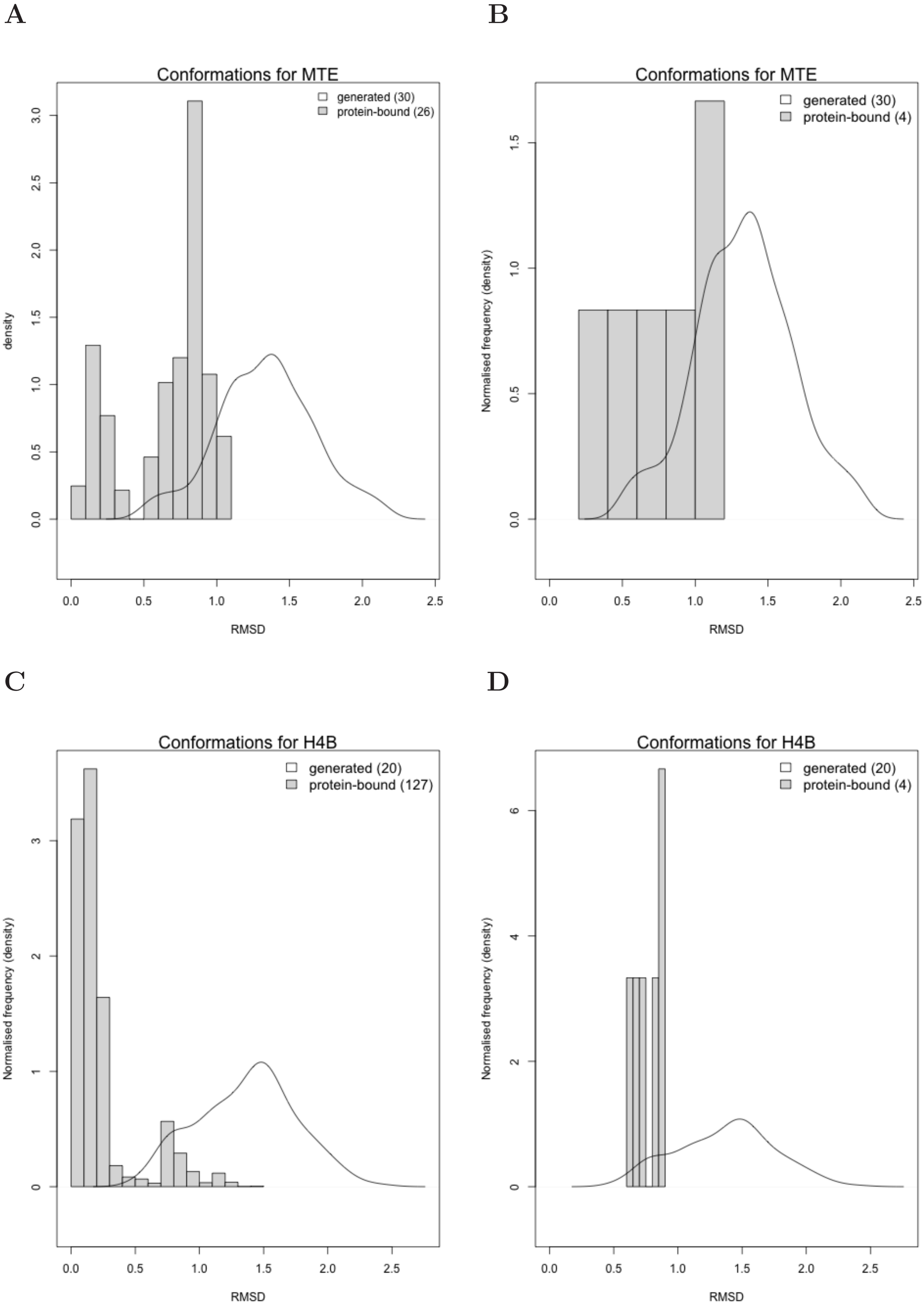


Figure 5.19: Density histograms and distributions of MTE (top) and H4B (bottom), based on the structures generated by OMEGA superimposed on the full data set (left) and the homology-filtered data set (right).

still shows both clusters but the sample size of 4 is too small to draw conclusions from the histogram.

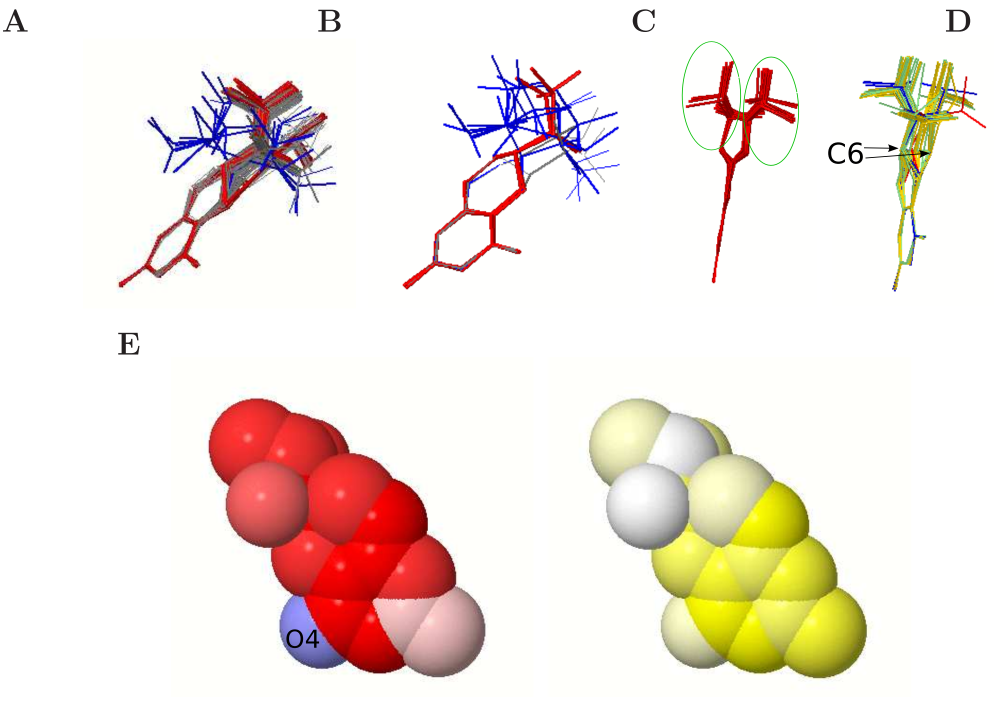


Figure 5.20: Conformations of H4B, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. All protein-bound (C) H4B structures showing the clusters (green circle) of the side chain conformers. Protein-bound structures coloured by CATH code (D). Average (left) and variance (right) in relative solvent accessibility of protein-bound H4B structures (E).

The homology analysis (see figure 5.20D) shows that the H4B structures of each CATH domain appear in both clusters, which implies that this conformational difference is independant of the superfamily. There are 66 protein-bound H4B structures with known CATH codes and their average pairwise RMSD is 0.17±0.23Å. 3.90.340.10 and 3.90.1230.10 are the two most-populated CATH families with average RMSD values of 0.17±0.13Å on 34 structures and 0.14±0.23Å on 29 structures, respectively. These values indicate that the cofactor conformation is not determined by the superfamily of the protein. Figure 5.20D confirms this conclusion, as a grouping by CATH code cannot be observed.

The solvent accessibility analysis shows that most of the cofactor is deeply buried in the protein, with the exception of the catalytically active O4 atom (see figure 5.20E).

5.2.2.6 General observations for the mono-nucleotide cofactors

All of the mono-nucleotide cofactors are more stretched out in the protein-bound than in the generated data. However, not all of the conformational variability appears to depend on the homologous superfamily of the protein domain, to which the cofactor is bound. While this homology-dependant conformational variability is not observed in COA, SAM and FMN show spatial clustering by CATH code, which indicates that the domain has an influence on the cofactor's conformation. While at least some atoms of COA and SAM are partly solvent accessible, the other mono-nucleotides are deeply buried in the binding domain. A tendency for catalytically active atoms to be more exposed than others is observed.

5.2.3 Group 3: large cofactors

Unfortunately, the 3D data set does not comprise any data for cofactors in group 3 of the 2D clustering. For B12, F43, SRM, HEC and HEA, OMEGA could not produce any conformations, since the force field does not support metal atoms. Leaving the metal atom out or replacing it by carbon or phosphorus resulted in chemically unrealistic conformers.

5.2.4 Group 4: dinucleotides

From structural group 4 (see figure 4.11D), this data set contains conformational variability conformations for FAD, NAD, NAP and MGD.

5.2.4.1 MGD

For MGD, figure 5.21A immediately reveals a large discrepancy between the generated and the protein-bound structures. The cofactor is a prime example for an extended conformation in the protein-bound data as opposed to much more compact conformations in the generated data set.

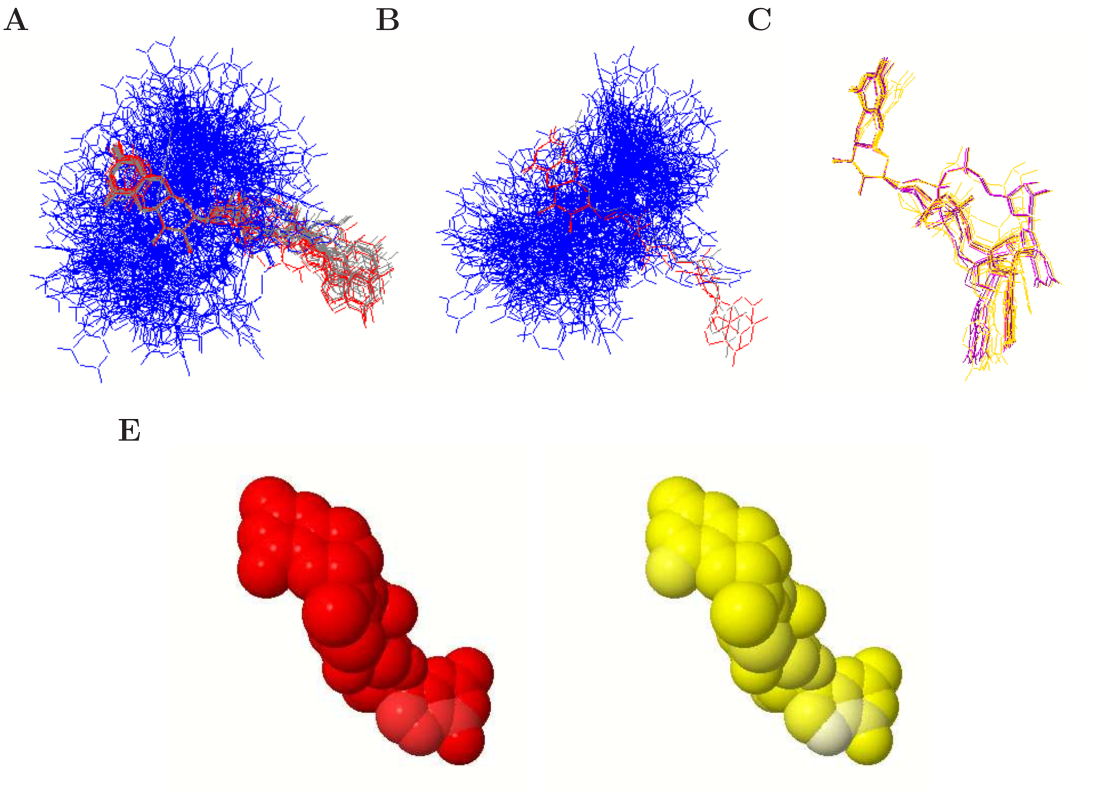


Figure 5.21: Conformations of MGD, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. Protein-bound structures coloured by CATH code (C). Average (left) and variance (right) in relative solvent accessibility of protein-bound MGD structures (D).

OMEGA does not generate the stretched out conformations at all, most likely due to high energy values above the cut-off. The distribution plot (see figure 5.22A) confirms this conclusion. While the average RMSD among the PDB structures peaks at 1Å, the much more varied conformers generated by OMEGA have an average RMSD of more than 3Å. The results from the homology filtering (see figures 5.22B and 5.21B) indicate that this holds true when redundant homologs are removed.

The homology analysis (see figure 5.21C) shows that, for this cofactor, the conformation does not appear to depend on the CATH family, as the average pairwise RMSD values of one of the CATH families (3.40.50.740, 3.39±1.38 on

6 structures) is even larger than the value calculated on all MGD structures (3.03±1.30 on 24 structures). However, these sample sizes are very small and this particular result may change when more structures become available.

MGD is deeply buried, as the solvent accessibility analysis in figure 5.21D shows. All atoms are less than 5% exposed to the solvent.

5.2.4.2 NAD and NAP

A similar picture arises from the data for NAD and NAP. The results are very similar for both molecules, and thus only the pictures for NAD will be shown in this section. The generated conformers are more compact and “folded back” onto the niacin ring, while the large majority of the PDB structures are extended. In fact, the compactness of the generated structures is so pronounced that it is very difficult to see the niacin ring in a 2D snapshot of the 3D applet (see figure 5.23A). When looking at a random sample of the PDB structures singly, one can observe that there are two major conformational clusters in NAD, depending on the angle and orientation between the niacin and the neighbouring ribose ring. These clusters correspond to the two peaks at 0.9 and 3.2Å in the histograms (see figures 5.24A and C). The extent of compactness of the whole molecule varies within both of these two conformational clusters. The two nucleotide-connecting phosphates appear to contribute most to this conformational variability within each cluster. The generated structures, too, vary the most at the two phosphate atoms. Although there is some overlap with the less compact end of the conformational spectrum of the PDB structures, the general tendency of higher compactness in the generated conformers is clearly visible (see figures 5.23A an B). The multimodal histogram for the full data set as opposed to the unimodal histogram for the CATH-filtered data set might indicate a homology-based bias in the full data set.

Looking at figures 5.23C and D, a spatial clustering by homology is indeed observed. The general average pairwise RMSD of all NAD structures is 10.05±2.47Å, based on 540 NAD structures that have assigned CATH codes. This is the highest value (with a relatively low standard deviation) seen among all cofactors in this data set. It is worth noting that NAD is one of the largest and

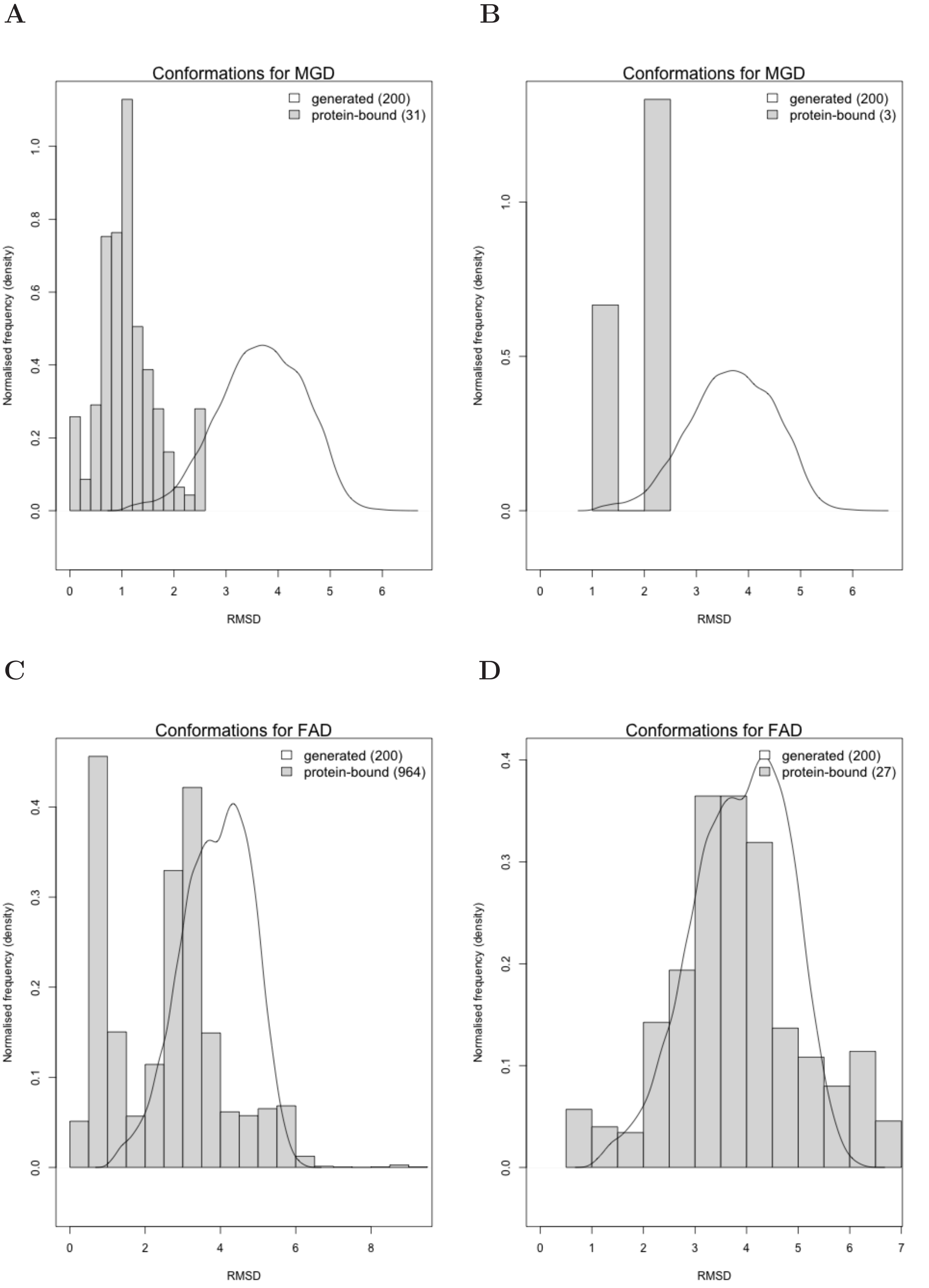


Figure 5.22: Density histograms and distributions of MGD (top) and FAD (bottom), based on the structures generated by OMEGA superimposed on the full data set (left) and the homology-filtered data set (right).

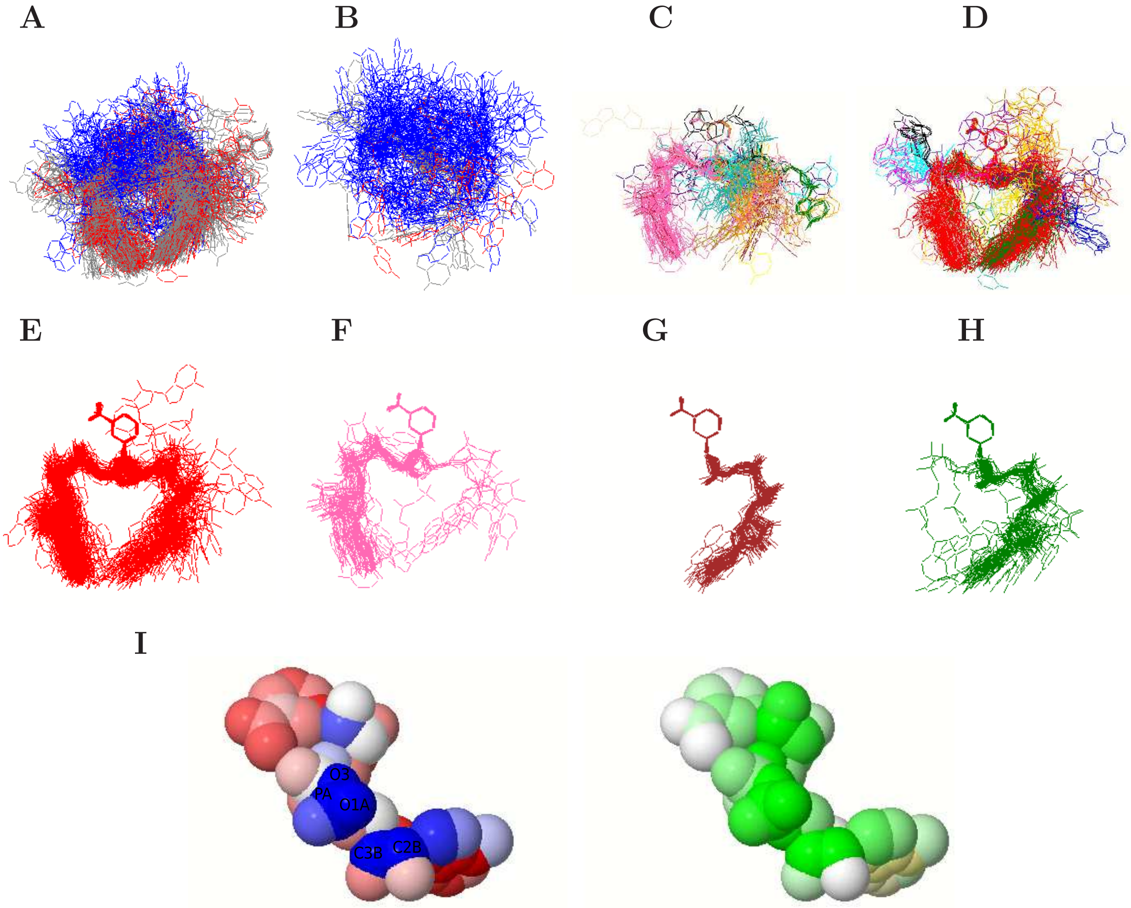


Figure 5.23: Conformations of NAD, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. Protein-bound structures that have CATH codes assigned coloured by CATH code (C,D). NAD structures bound in CATH families 3.40.50.720 (E, red), 3.30.360.10 (F, pink), 3.90.180.10 (G, maroon) and 3.90.110.10 (H, green). Average and variance in relative solvent accessibility of protein-bound NAD structures (I).

most flexible cofactors in the data set, which may contribute to this high average RMSD value. The most populated CATH family is 3.40.50.720 (which is unsurprisingly the so called "NAD(P)-binding Rossmann-like Domain"). With 240 structures, this family contributes 44% to the full data set. Figure 5.23E shows a superposition of all these NAD structures. Their average pairwise RMSD is comparably high with a value of 8.31±1.69Å. The next most-populated domain family (figure 5.23F, 47 structures) is CATH node 3.30.360.10 with 47 structures, which is mostly, but not exclusively occupying only one of the two spatial clusters. The average pairwise RMSD is 9.93±2.44Å, comparable with the previous family's. The results are similar for CATH node 3.90.110.10 (figure 5.23H) with 8.75±1.48Å at a sample size of 37 structures. CATH family is 3.90.180.10 also comprises 37 NAD structures, which are exclusively located in one of the two clusters. Therefore the average pairwise RMSD is significantly lower with 1.50±0.97Å. The results show that the conformation of this cofactor does depend on the homologous superfamily of the binding protein, in spite of the high average pairwise RMSD values, which are caused by the two-cluster trend in most of the families.

The solvent accessibility analysis shows that the nicotinamide ring and most of the adenine ring are buried. Some atoms in the phosphate and ribose portions have a relative solvent accessibility of $\geq 25\%$, but those also coincide with a high variance, indicating that they are only buried in some of the enzymes (see figure 5.23I). Overall, NAD (and NAP) is highly buried and its conformation appears to be largely determined by the homologous superfamily to which it is bound.

5.2.4.3 FAD

Lastly, FAD also shows a high separation between the generated and protein-bound structures. The same trend as in the other dinucleotides is observed: stretched out PDB conformers and compact generated conformers (see figures 5.25A and B). Similar to the observations for NAD and FMN, two clusters of conformations exist for both the generated and the observed structures, depending on the angle that the C1'-C2' bond forms with the flavin plane. The bimodal histogram documents this fact for the protein-bound structures (see figure 5.22C).

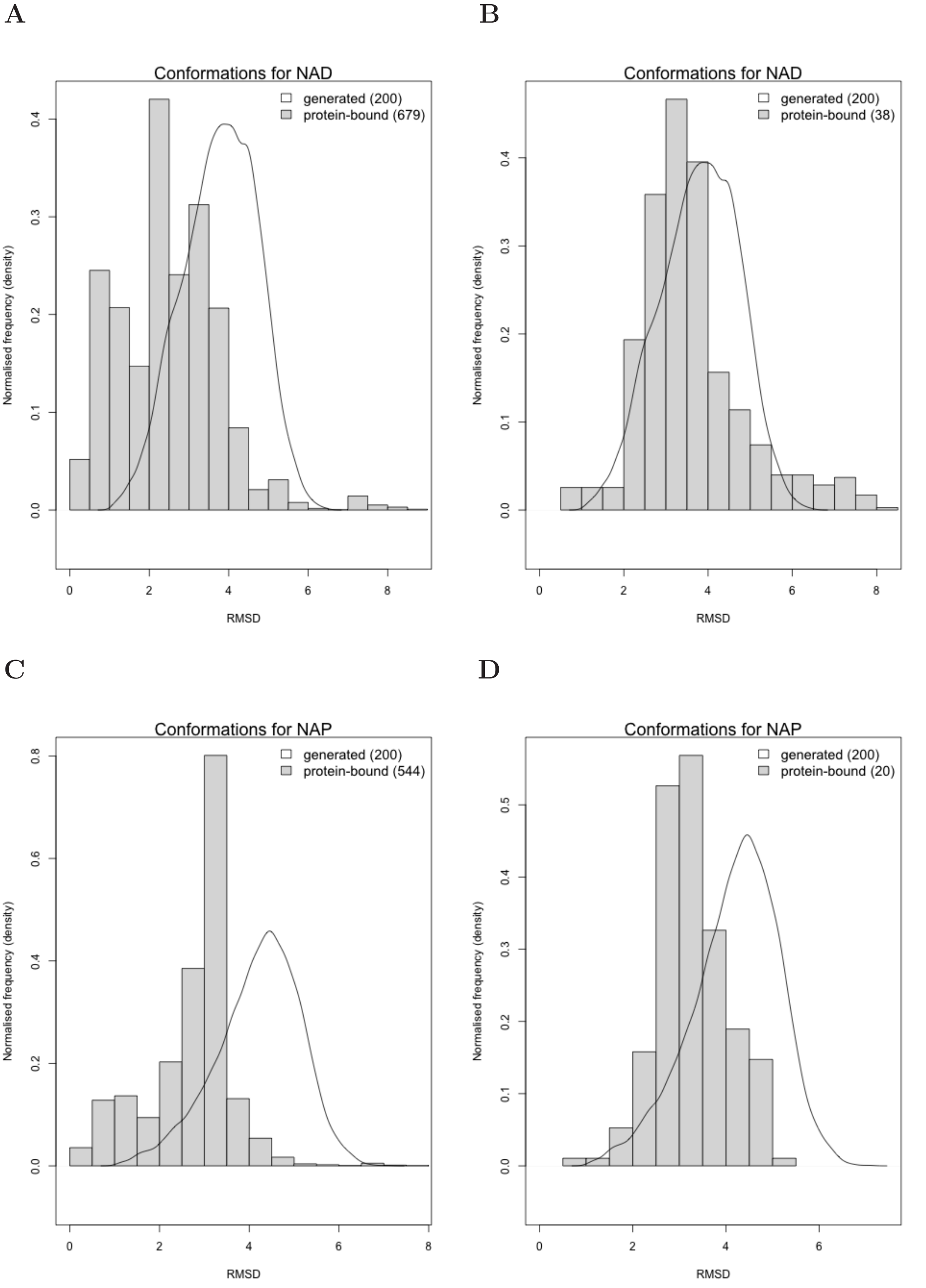


Figure 5.24: Density histograms and distributions of NAD (top) and NAP (bottom), based on the structures generated by OMEGA superimposed on the full data set (left) and the homology-filtered data set (right).

The trend is not as obvious for the generated structures as can be seen in the distribution graph. The loss of the first histogram peak upon CATH-filtering (see figure 5.22D) might indicate a bias introduced by a high number of crystal structures of the same or very similar proteins.

Figures 5.25C and D show the superposition of all those protein-bound FAD structures with CATH codes. Overall there are 606 such structures with an average pairwise RMSD of 8.15±2.48Å. The 5 most populated CATH families are shown in figures 5.25E-I. The most populated family identified in mint by CATH code 3.50.50.60 has an average pairwise RMSD of 3.37±3.78Å. This homologous superfamily comprises 227 structures (corresponding to over one third of all FAD structures). The conformation of these structures is clearly more restricted than the overall FAD conformation. At the topology level of the CATH classification, the 3.50.50.x topology is labelled "FAD/NAD(P)-binding domain". The next four most populated families are 3.30.390.30 with 3.22±3.83Å (53 structures), 3.40.50.80 with 3.35±2.23Å (43 structures), 2.40.30.10 with 4.71±1.57Å (35 structures) and 3.90.660.10 with 1.31±0.50Å (34 structures). All of the FAD cofactor conformations in these homologous superfamilies are clearly more conserved than the overall conformational space adopted in the protein-bound data set overall.

The solvent accessibility analysis shows similar results as for FMN and NAD, with a generally largely buried cofactor (see figure 5.25J).

5.2.4.4 General observations for the dinucleotide cofactors

The difference between the conformational variability between the protein-bound and the generated cofactor conformations is most pronounced for the dinucleotide cofactors. In all examples, the protein-bound cofactors are more extended, whereas the ones generated by OMEGA, to simulate the conditions in solution, are much more variable. For FAD and NAD, the homology analyses showed that the conformation that the molecule adopts in the protein depends partially on the identity of the homologous superfamily (identified by its CATH code). For MGD, this trend was not observed, which may be a side effect of a too small data set. It is

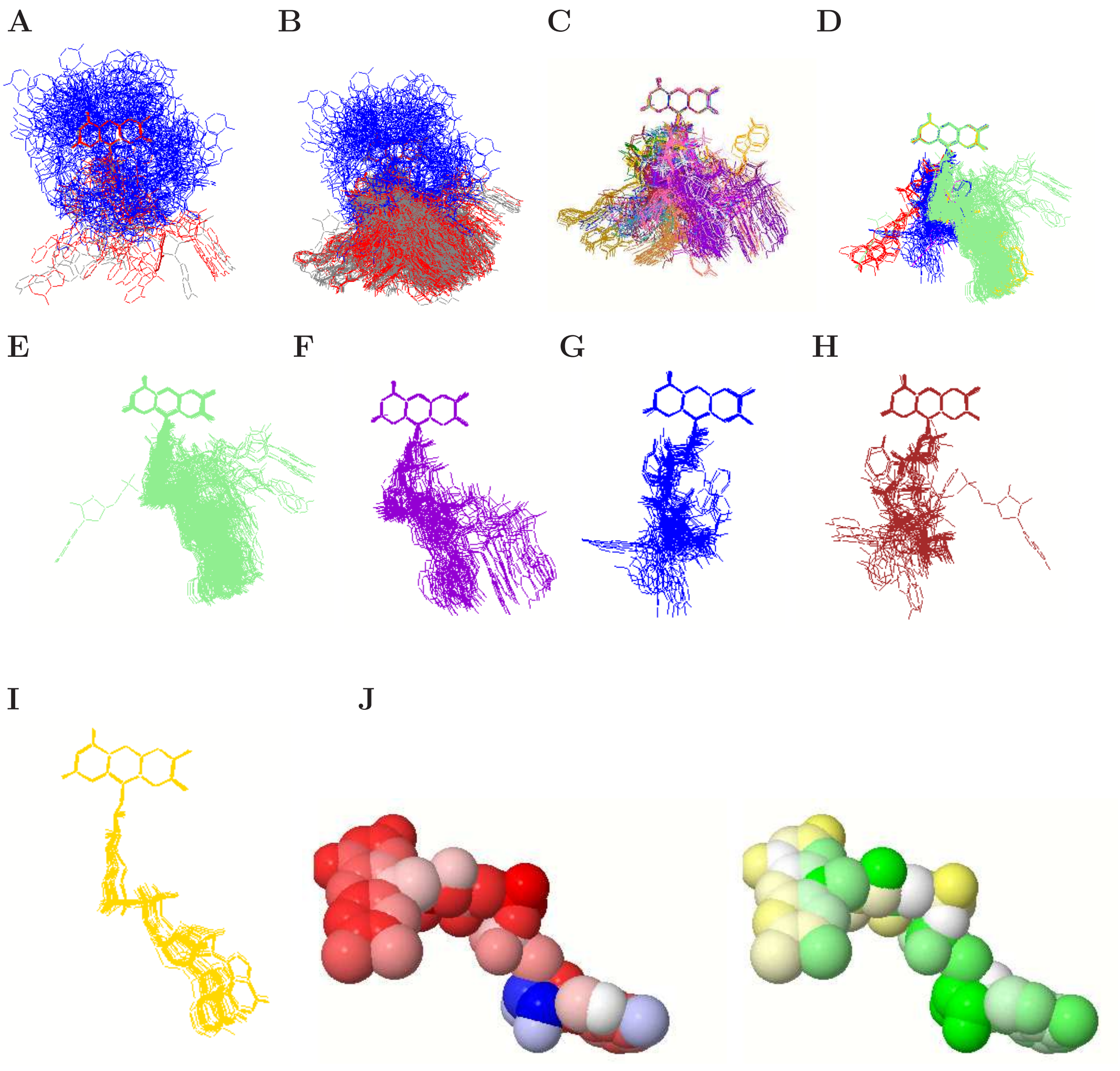


Figure 5.25: Conformations of FAD, protein-bound (red ≤2Å, grey >2Å) and computer-generated (blue). A superposition of the full (A) and CATH-filtered (B) data set onto the generated structures. Protein-bound structures that have CATH codes assigned coloured by CATH code (C,D). FAD structures bound in CATH families 3.50.50.60 (E, mint), 3.30.390.30 (F, purple), 3.40.50.80 (G, red), 2.40.30.10 (H, maroon) and 3.90.660.10 (I, yellow). Average and variance in relative solvent accessibility of protein-bound FAD structures (J).

further observed that MGD is clearly less variable in the protein-bound data set than NAD and FAD are.

In all examples, most of the atoms are on average highly buried. In the case of MGD, this is true for the whole molecule and for FAD most of the atoms are buried. For NAD, the two ring systems, are less exposed to the solvent than some of the atoms of the phosphate and ribose moieties.

5.3 Discussion

5.3.1 Limitations of the method and the data set

Although the PDB database has been growing quickly over recent decades and even months, the amount of experimentally determined ligand structures is still sparse, especially when the conformational variability of specific small molecules is to be investigated. Some cofactors, like NAD, FAD, PLP and Heme appear in several hundred crystal structures, but for most of cofactors, less than 100 structures are available. The whole spectrum of PDB entries cannot be expected to be a non-redundant representation of protein fold space or even enzyme space, for the following reasons: first, there is most likely a bias in PDB entries towards those structures that are easy (or even possible) to crystallise; second, those proteins that are of importance to human diseases or to basic cellular metabolism are more likely to have been crystallised than others; and third, some proteins' structure has been documented with several crystal structures (for example to crystallise different steps in the reaction of an enzyme or simply in order to get higher resolutions of the same important protein using better technology). All of these biases imply that the already sparse data set should be filtered to avoid redundancy and thus a bias in the results.

The methods used have some caveats, too. As mentioned above, the experimental data for various conformations of the same cofactor in solution is sparse and the usage of a conformation generation program is merely a prediction (albeit not a bad one). Although Chen & Foloppe (2011) conclude, that conformational sampling for drug-like molecules is essentially a solved problem, this might not necessarily be true for all protein ligands. The fact that all of the force fields that are integrated with OMEGA cannot be used to generate conformations of metal-containing molecules, is a problem here, too. The superposition program used has been found to sometimes produce wrong superpositions, depending on which atoms are chosen for the RMSD minimisation. If there are several solutions to the superposition, the procedure cannot distinguish between the right and the wrong one, and the visualisation wrongly suggests that some of the molecules are the wrong stereoisomer. This problem was only detected at a very late stage in this work, and thus some of the superposition visualisations (e.g. for MTE or BTN) in the CoFactor database need to be corrected in the next release. Alternatively, a superposition program could be developed for the specific task here: instead of matching a set of atoms from one structure to a set of atoms in another one (thus risking the possibility of more than one match), the new program could take two ordered lists of atoms as input, in which the same index in the two lists indicates equivalent atoms for the superposition. Of course, this may cause other problems, for example with the exact matching of, say, the four oxygen atoms bound to a phosphorus atom. In the ordered list approach, the superposition program needs to know exactly which of the four O atoms in the one structure corresponds to which one in the other structure, or the result is wrong yet again. However, this assignment is often not possible as the oxygen atoms are equivalent.

If this project was started again on a big enough data set, it may be worth investigating if there are better methods than calculating the average and standard deviation RMSD to mathematically investigate the significance of conserved ligand conformations in homologous proteins.

Finally, there are the general limitations to the accuracy of the coordinates of ligands in crystal structures, as discussed in the Introduction. The B-factors and the correlation coefficient could be used to exclude less accurate data and programs like CCDC's MOGUL could be applied to check whether the bond lengths and bond angles are chemically realistic. Of course all of this would be very time-consuming and further reduce the already small data sets.

5.3.2 Conclusions

In general, this study confirms the findings of Stockwell & Thornton (2006), namely that cofactors are - much like the three ligands in the original work - less compact and more extended when they are bound to a protein. While Stockwell & Thornton (2006) have compared the protein-bound ligand structures to a representative sample of the whole, theoretically possible conformational space of that molecule, in this work they are compared to the low-energy conformations in solution, as calculated by OMEGA (OpenEye Scientific Software, 2004–2010). The main difference is that in OMEGA uses an energy cut-off to distinguish likely conformations from unlikely ones.

The results in this chapter indicate that the observed effect (for cofactors to be more extended when bound to the protein) depends on the size and flexibility of the cofactor, as larger molecules such as dinucleotides show the effect to a greater extent than the medium-sized mono-nucleotides and the small cofactors. This is expected in that smaller molecules have on average fewer rotatable bonds, thus exhibit lower flexibility, and therefore show lower pairwise RMSDs. The mono-nucleotides and small cofactors contain molecules that mainly consist of ring systems, such as MTE, H4B, ASC and especially PQQ. Due to the lower flexibility of these cofactors the difference between the extended (protein-bound) and compact (generated) conformations is reduced or in the case of PQQ non-existent. Further, it is worth noting that some cofactors are the exception to this rule, such as COA, whose conformational variability is clearly higher than other cofactors.

The reason for the more extended conformation is possibly to allow the substrates to access the catalytically active portions of the cofactor. If the cofactor was bound in a way that the non-active portions had the possibility to obstruct access to the cofactor, catalysis would occur less efficiently or possibly not at all. The larger (and more flexible) the non-catalytic portion, the more likely is this problem.

The question whether the conformation of cofactors is similar in homologous proteins has also been investigated. The results suggest that, for many superfamilies, the conformation of the cofactors is conserved in all relatives. However, there are some clear exceptions in the results for TPP, COA, H4B and MGD.

The solvent accessibility results show that most of the cofactors are largely buried in the biological assembly. Often, the catalytic atoms are more exposed to the solvent than other cofactor atoms, possibly to allow substrate accessibility. Phosphate groups are more exposed in general, as they are highly flexible and charged.

These results have implications for the use of homology information in docking procedures, which are often used in computational binding studies, e.g. in drug development. The conformation of a ligand from an experimentally determined protein structure is often taken as the starting conformation for ligand-protein docking. The results above indicated that this may only be useful if the known protein structure is a homolog of the one in the docking procedure, as for most of the data set, the conformation of the cofactor was conserved within the homologous superfamily. However, even this can not always be assumed. In the case of TPP, for instance, the cofactor adopted two different conformations for the same CATH code. Some ligands, such as COA, are extraordinarily flexible and homology information does not provide a complete solution in the docking task. Unfortunately the results suggest that the usefulness of using a specific starting conformation from another structure must be evaluated on a case-by-case basis. Generally, homologous proteins are clearly more likely to bind the same ligand in a similar conformation, and mono-nucleotides as well as dinucleotides appear to be more extended than solution-based conformation generation programs like OMEGA might suggest. This study also shows that conformation generation programs are not necessarily the best way to choose a starting conformation, although at least a subset of the generated conformations was mostly observed in experimentally determined structures.

To sum up it may be concluded that neither the experimental data, nor the methods available suffice to draw much stronger conclusions than the one published by Stockwell & Thornton (2006). However, this work has shown that their conclusions are not necessarily applicable for all ligands and that this topic should be revisited when more data are available.

Chapter 6

Conclusions

While organic enzyme cofactors have long been known to play a crucial role in enzyme catalysis, in this work these molecules have been analysed in more detail as one of the three catalytic entities in enzymes. Although generating a precise definition of what constitutes an organic enzyme cofactor is difficult, a definition, based on criteria as objective as possible while still conforming with the prevalent ideas, has been developed. Furthermore, the relevant biological, chemical and crystallographic background as well as the existing resources in the broader field, have been summarised.

6.1 Summary of the findings

The CoFactor database was developed as a resource offering a comprehensive entry point for scientists interested in one of the organic enzyme cofactors. It provides manually curated information from relevant scientific literature on key facts about the molecule, its molecular function as a cofactor, its chemical structure and the enzyme mechanisms, in which it is involved. CoFactor further integrates cofactor-related information about enzyme reactions, protein sequences, biological species, three-dimensional protein structures and structural protein domains, and it includes powerful visual representations of enzyme space, conformational space and solvent accessibility. The relative solvent accessibility visualisations and superpositions of the observed conformations offer insights into the binding mode of each cofactor and provide the basis for the more detailed conformational variability analysis performed.

The CoFactor database was used as the basis for answering several important questions. A principal component analysis of the physicochemical properties of the cofactors revealed that the cofactors mainly vary in polarity and size, followed by flexibility as the third-largest contributor. A comparison to the physicochemical properties of other metabolites in the cell showed that cofactors are not restricted to a narrow subset of the chosen property descriptors, but instead are sampled from most of metabolite-property space, with the exception of a small subset of that space, which is populated by medium-sized and more hydrophobic molecules. This suggests that organic cofactors are not easily distinguishable from other metabolites based on the physicochemical properties only. The comparison of the cofactor and metabolite variable distributions showed that their means are only significantly different in the polarity-measuring variables.

The question of the existence of intrinsic groupings in the cofactor data set has been investigated on three levels. First, the clustering analysis based on the physicochemical properties showed agreement with the principal component analysis. Second, the clustering analysis based on the pairwise similarity of the cofactor data set's two-dimensional chemical structures started to document common structural motifs. Four general groups of cofactors were thus identified: mononucleotides, dinucleotides, large cofactors and small cofactors. The large cofactors may be further subdivided into three groups, two of which have an obvious common structural motif (porphyrin-like cofactors and large quinones with long hydrophobic side chains). The group of small cofactors is more diverse; a common motif is not detectable, most probably due to the small size of the molecules, although most of the molecules in this group comprise at least one ring. Thirdly, the grouping question was addressed by a manual classification of the cofactors based on the overall reactions the cofactors co-catalyse. The resulting classification is non-hierarchical and comprises two major and five minor functional classes. The two major ones are the redox cofactors and group transfer cofactors.

A functional analysis of the cofactor data set was undertaken. The overall reaction level of investigation showed that there is an over-representation of organic cofactors in the oxidoreductases and an under-representation in the hydrolases. The former may be explained by the unique hydride transfer ability of some organic cofactors (among other cofactors from the redox class), while the latter may be a consequence of the relative simplicity of the reactions of hydrolases and their use of magnesium (and thus a metal cofactor) activated water molecules. The analysis on the more detailed mechanistic level was based on the MACiE database and revealed that organic enzyme cofactors perform eight of the nine groups of mechanistic function. The most striking result from this analysis was that neither metal cofactors nor amino acids perform the hydride shuttle function in the MACiE data set. This indicates that this function is unique to organic cofactors and thus complements the catalytic toolkit.

The analysis of the conformational variability and solvent accessibility of organic enzyme cofactors revealed that previously published conclusions (that ligands are more extended when bound to the protein than expected when compared to all theoretically feasible conformations) hold true for groups of cofactors to varying degrees. The dinucleotide cofactors confirm these findings but there are some exceptions, mostly among the mono-nucleotide and small cofactors. It is worth noting that some aspects of the different analyses in this work integrate well. The structural groups, especially the dinucleotide cofactors, determined based on the chemical two-dimensional structures of the cofactor molecules mostly correlate with the degree of variability in conformation.

In spite of the successful investigation of the questions and the findings described above, some caveats remain. Although the PDB database keeps growing, and more and more structural data is available to the community, the number of high-resolution structures of ligands bound to proteins is often still too small for statistically relevant conclusions, especially when a data set is required not to include evolutionarily related proteins. The pipeline used for the conformational variability analysis further reduces the data set, for example by the conformation generation program, whose force field does not support metal atoms. At a late stage of this analysis, a bug in the superposition software was found which requires that some of the superposition data in the CoFactor database be corrected. This is planned for the next release of the database, together with some minor mistakes that were pointed out by users.

6.2 Future work

In addition to the aforementioned update of the CoFactor database the aim is to add query functionality to the database web pages and to update the literature-based information. The next version of CoFactor will additionally include a much better coverage of structural information, as it will be taken directly from PDBsum rather than from the discontinued PROCOGNATE database. There are plans at the EBI to integrate all of the enzyme-related data resources into an enzyme portal, which will offer a unique entry point for researchers seeking information about enzymes.

A group-internal collaboration is planned, in which the evolution and distribution of cofactor-binding enzyme domains shall be studied. The FunTree project is based on the FunTree database (Dr. N. Furnham, unpublished), which uniquely combines phylogenetic information of enzyme domains with information about their overall reaction similarities, small molecule similarities of their substrates and products and three-dimensional structural similarity. The objective is to investigate the relationship between domain structure (homologous superfamily) and enzyme function. How many homologous superfamilies utilise one cofactor and what is each family's functional repertoire? For each of these families: how were the different functions obtained in the course of evolution and how do interacting domains relate to each other? How similar are their overall reactions and small molecules involved? Answering these and related questions will widen the understanding of the contributions of structural and functional constraints and requirements to the course of enzyme evolution.

The question whether (or to what extent) the catalytic power of a cofactor-dependant enzyme is due to the cofactor or the enzyme is worth investigating. Of course, some enzymes will crucially depend on both, but examples are possible where either or both entities may be able to catalyse the reaction on their own, but possibly with a reduced efficiency. In order to obtain a homogeneous data set, this project would have to involve a collaboration with experimentalist, who could measure catalytic rates of a large set of cofactor-dependent enzymes under similar experimental conditions.

6.3 Reflections on the role of cofactors in early evolution

Organic enzyme cofactors are also fascinating molecules in the context of early evolution. It has been found that most cofactors are made up of a set of building blocks. The comparative PCA has shown that the chosen physicochemical descriptors may be used to find groupings within the cofactors, but cannot be used to distinguish catalytic molecules from other metabolites. The structural scaffold of the cofactor molecules, however, is in some cases (porphyrins and quinones) cofactor-specific. Thus, some of these building blocks are common metabolites, like nucleic acids, amino acids and fatty acids, whereas others appear to be cofactor-specific molecules whose unique structure enables them to add functions to the catalytic toolkit available through amino acids and metal ions alone. Assuming that the RNA world hypothesis for early evolution is relevant, thinking about the role of these cofactors inspires interesting questions and scenarios. Why has the rather small number of cofactors emerged? What is their species distribution? At what stage did eukaryotes lose the ability to synthesise many cofactors? Are cofactor binding sites good targets for drug design? Why is it that the combination of a phosphate group, a ribose group and a base is so successful in evolution, given that nucleotides not only encode (DNA) and transfer (mRNA) genetic information, but also catalyse (ribozymes) chemical reactions and widen the catalytic repertoire (organic cofactors)? Many theories and models have been published on several aspects of the RNA world and possible transition scenarios to today's cells, in which proteins act as the main catalytic entity. Although the scope (and time available) of this work does not extend to deeper analyses and considerations in the matter, the author believes that organic cofactors may have played a crucial role in the RNA world, the protein world and the transition. It is even questionable if today's cellular catalytic world should be called a "protein world", given that non-amino acid molecules like catalytic RNAs, metal ions and organic cofactors still contribute crucially to the essential catalytic reactions we observe now.

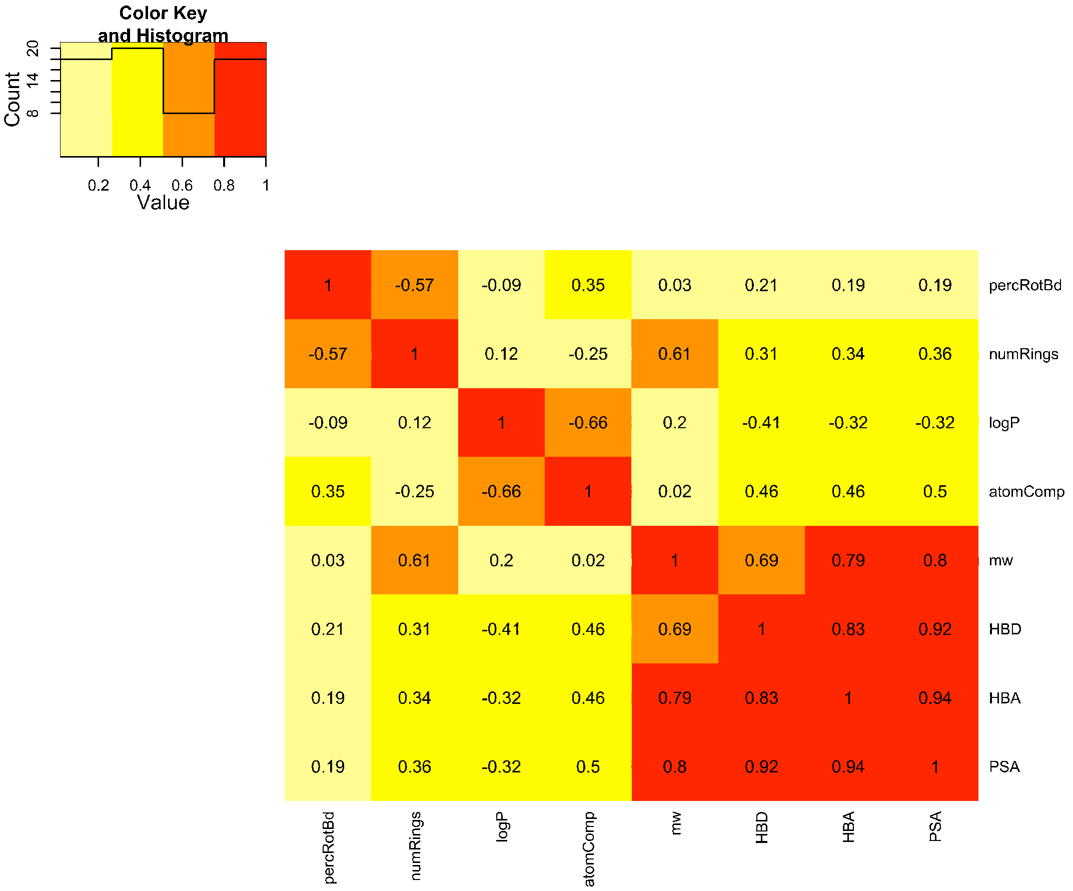
Although developing computational approaches to investigate these topics further are difficult to design (due to lack of phylogenetic information from the RNA world), the topic is hugely fascinating. One potentially interesting project would be to estimate the relative ages of the cofactors, *i.e.* the relative time point at which the genetic information to synthesise it became available in an organism's genome. Because the biosynthesis of some cofactors depends on enzymes using other cofactors, one could make a "dependency map" of these biosynthetic relationships, using the manually curated biosynthesis information in the CoFactor database as a starting point. The degree of centrality in the metabolic network of the enzyme reactions catalysed by each cofactor, combined with the knowledge about the functional profile overlaps between the cofactors, could give an estimate of the functional uniqueness and metabolic importance of the cofactor, which may also be an indicator for its relative age. However these ideas are highly speculative.

In general, this project has shown (as many have before) that classification in biology is tricky and sometimes subjective. For the human brain, classifying objects into manageable categories is useful in order to simplify the complex relationships between a plethora of objects in the world. Biology, however, is different. It has evolved from a set of molecules, which arose from the elements and conditions on an early Earth. The only constraints in evolution were the availability of the elements and molecules, and their usefulness for self-replicating organisms. Cofactors do not fit easily into the current way of classifying molecules of biological relevance. They are not storing information like DNA does and they are not transferring information like mRNA does. They do transfer chemical moieties, from single electrons to whole portions of molecules, and they do take part in catalysis, like proteins do. In this work it has been shown that although biology and biochemistry is complex, classification is possible and often useful. Although defining what a cofactor is exactly may still be subjective in some cases, the author (along with many other scientists) has found it useful to look at organic cofactors as a category of molecules, namely as one of the three entities in enzyme catalysis.

One of the main objectives of this work was to formalise and organise the knowledge about organic enzyme cofactors as one of these three biocatalytic entities. The findings in this thesis and the CoFactor database achieve this. CoFactor complements the knowledge about the other two biocatalytic entities, namely catalytic amino acids (MACiE database) and metal ions (Metal MACiE database). These three resources together offer easy access to insights about many aspects of the catalytic mechanisms of enzymes.

Appendix A

Additional figures



	percRotBd	numRings	logP	atomComp	mw	HBD	HBA	PSA
percRotBd	1	-0.57	-0.09	0.35	0.03	0.21	0.19	0.19
numRings	-0.57	1	0.12	-0.25	0.61	0.31	0.34	0.36
logP	-0.09	0.12	1	-0.66	0.2	-0.41	-0.32	-0.32
atomComp	0.35	-0.25	-0.66	1	0.02	0.46	0.46	0.5
mw	0.03	0.61	0.2	0.02	1	0.69	0.79	0.8
HBD	0.21	0.31	-0.41	0.46	0.69	1	0.83	0.92
HBA	0.19	0.34	-0.32	0.46	0.79	0.83	1	0.94
PSA	0.19	0.36	-0.32	0.5	0.8	0.92	0.94	1

Figure A.1: Correlation matrix of the metabolite PCA (see section 4.1.4) from the metabolite data set (described in section 2.2.1.2)

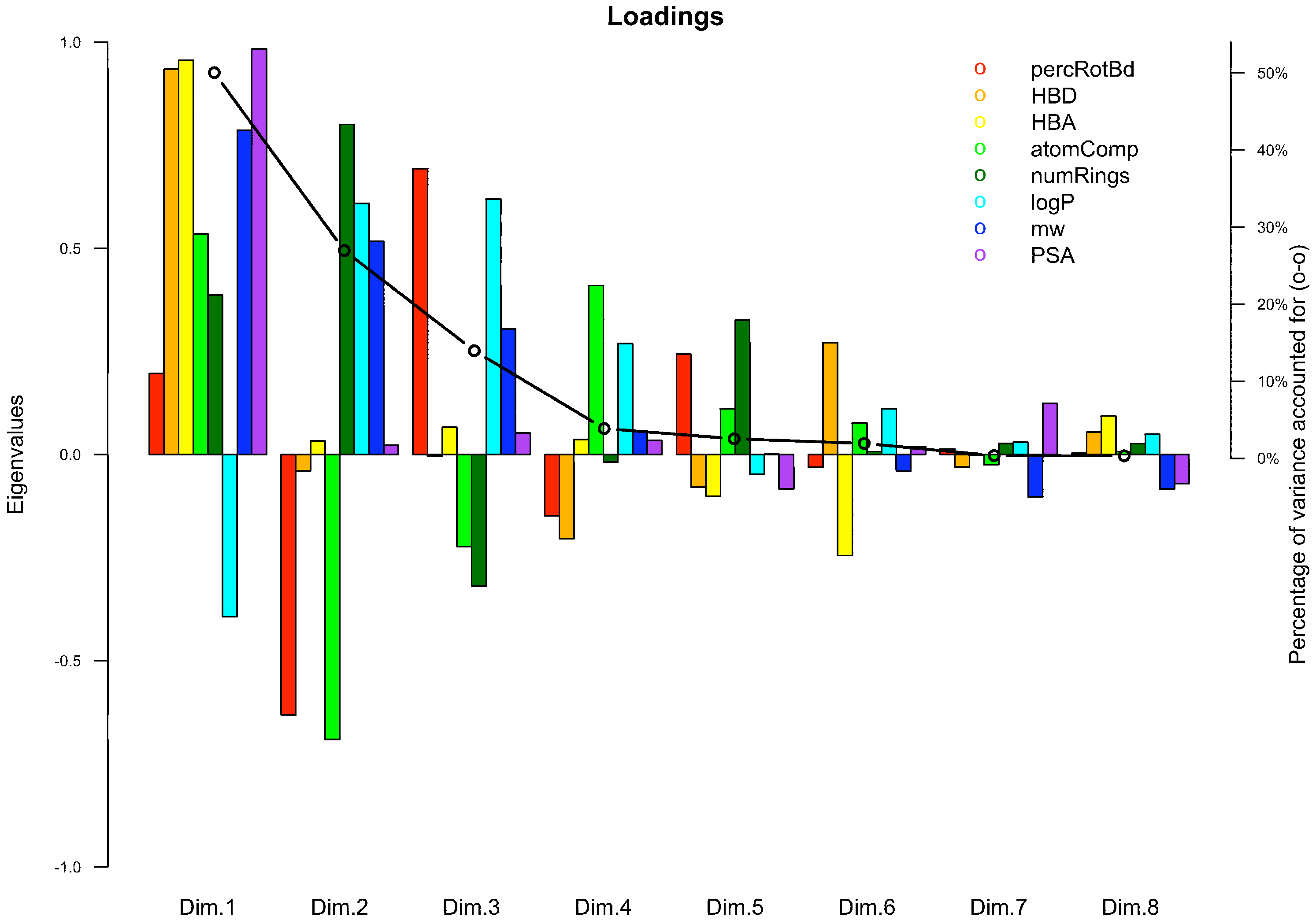


Figure A.2: Loadings of the metabolite PCA (see section 4.1.4) from the metabolite data set (described in section 2.2.1.2)

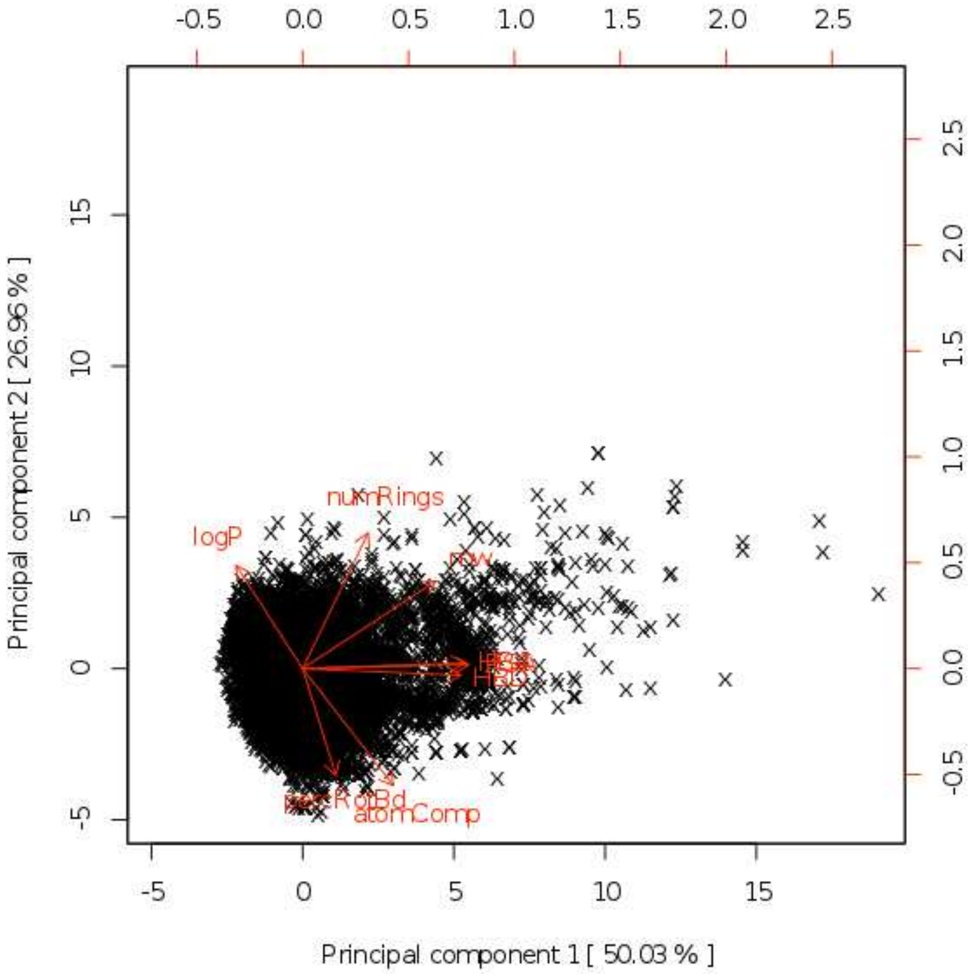


Figure A.3: Biplot of the metabolite PCA (see section 4.1.4) on the metabolite data set (described in section 2.2.1.2)

References

Abdi, H. & Williams, L.J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews Computational Statistics*, **2**, 433–459. 45

Allen, F.H. (2002). The cambridge structural database: a quarter of a million crystal structures and rising. *Acta Crystallogr B*, **58**, 380–8. 154

Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C. & Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*, **36**, D419–25. 3

Andreini, C., Bertini, I., Cavallaro, G., Holliday, G.L. & Thornton, J.M. (2009). Metal-macie: a database of metals involved in biological catalysis. *Bioinformatics*. 26, 32, 58, 145

Aoyama, K., Watabe, M. & Nakaki, T. (2008). Regulation of neuronal glutathione synthesis. *J Pharmacol Sci*, **108**, 227–38. 80, 81, 82

Arroyo, A., Rodríguez-Aguilera, J.C., Santos-Oca texttildelowna, C., Villalba, J.M. & Navas, P. (2004). Stabilization of extracellular ascorbate mediated by coenzyme Q transmembrane electron transport. *Methods Enzymol*, **378**, 207–17. 88, 96, 97

Arun, K.S., Huang, T.S. & Blostein, S.D. (1987). Least-squares fitting of 2 3-D point sets. *IEEE Trans Pattern Anal Mach Intell*, **9**, 699–700. 55

Ashburner, M. (1998). Speculations on the subject of alcohol dehydrogenase and its properties in drosophila and other flies. *Bioessays*, **20**, 949–54. 150

Banik, R.M. & Prakash, M. (2004). Laundry detergent compatibility of the alkaline protease from bacillus cereus. *Microbiol Res*, **159**, 135–40. 21

Bartlett, G.J., Porter, C.T., Borkakoti, N. & Thornton, J.M. (2002a). Analysis of catalytic residues in enzyme active sites. *J Mol Biol.*, **324**, 105–21. 4

Bartlett, G.J., Porter, C.T., Borkakoti, N. & Thornton, J.M. (2002b). Analysis of catalytic residues in enzyme active sites. *J Mol Biol*, **324**, D129–133. 37

Bashton, M., Nobeli, I. & Thornton, J.M.M. (2006). Cognate ligand domain mapping for enzymes. *J Mol Biol*, **364**, 836–52. 33, 59

Bender, A. (2010). Databases: Compound bioactivities go public. *Nature Chemical Biology*, **6**, 309–309. 25

Benkovic, S.J. & Hammes-Schiffer, S. (2003). A perspective on enzyme catalysis. *Science*, **301**, 1196–202. 20

Benner, S.A., Ellington, A.D. & Tauer, A. (1989). Modern metabolism as a palimpsest of the RNA world. *Proceedings of the National Academy of Sciences*, **86**, 7054–7058. 86

Bentinger, M., Brismar, K. & Dallner, G. (2007). The antioxidant role of coenzyme Q. *Mitochondrion*, **7 Suppl**, S41–50. 90

Berger, B. & Leighton, T. (1998). Protein folding in the hydrophobic-hydrophilic (hp) model is NP-complete. *J Comput Biol*, **5**, 27–40. 27

Berkner, K.L. (2008). Vitamin K-dependent carboxylation. *Vitam Horm*, **78**, 131–56. 116

Berman, H., Henrick, K., Nakamura, H. & Markley, J.L. (2007). The worldwide protein data bank (wwpdb): ensuring a single, uniform archive of pdb data. *Nucleic Acids Res.*, **35**, D301–3. 24

Blakeley, M.P., Mitschler, A., Hazemann, I., Meilleur, F., Myles, D.A. & Podjarny, A. (2006). Comparison of hydrogen determination with x-ray and neutron crystallography in a human aldose reductase-inhibitor complex. *Eur Biophys J*, **35**, 577–83. 31

Bolton, E.E., Wang, Y., Thiessen, P.A. & Bryant, S.H. (2008). *PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 IN Annual Reports in Computational Chemistry.*, vol. 4. Elsevier. 23

Booth, S.L. & Al Rajabi, A. (2008). Determinants of vitamin K status in humans. *Vitam Horm*, **78**, 1–22. 89

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. (2007). UniProtKB/Swiss-Prot. *Methods Mol Biol*, **406**, 89–112. 33, 58, 65

Boutselakis, H., Dimitropoulos, D., Fillon, J., Golovin, A., Henrick, K., Hussain, A., Ionides, J., John, M., Keller, P.A., Krissinel, E., McNeil, P., Naim, A., Newman, R., Oldfield, T., Pineda, J., Rachedi, A., Copeland, J., Sitnov, A., Sobhany, S., Suarez-Uruena, A., Swaminathan, J., Tagari, M., Tate, J., Tromm, S., Velankar, S. & Vranken, W. (2003). E-msd: the european bioinformatics institute macromolecular structure database. *Nucleic acids research*, **31**, 458–462. 24, 38, 40, 59, 65

Branden, C. & Tooze, J. (1991). *Introduction to protein structure.*. Garland Publishing Inc., New York and London, 1st edn. 3

Bremer, M. & Doerge, R.W. (2010). *Statistics at the bench: a step-by-step handbook for biologists*. Cold Spring Harbour Laboratory Press. 44

Brunger, A. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*, **355**, 472–5. 29

Brunger, A.T., Leahy, D.J., Hynes, T.R. & Fox, R.O. (1991). 2.9 åresolution structure of an anti-dinitrophenyl-spin-label monoclonal antibody fab fragment with bound hapten. *J Mol Biol*, **221**, 239–56. 30

BUCKEL, W. & GOLDING, B.T. (2006a). Radical enzymes in anaerobes. *Annual review of microbiology*, **60**, 27–49. 73, 86

BUCKEL, W. & GOLDING, B.T. (2006b). Radical enzymes in anaerobes. *Annu Rev Microbiol*, **60**, 27–49. 90, 96, 97, 98

BUGG, T.D.H. (2004). *Introduction to enzyme and coenzyme chemistry.*. Blackwell publishing Ltd, Oxford, 2nd edn. 2, 4, 5

CATH (2009). CATH glossary: domain. `http://www.cathdb.info/wiki/doku.php?id=glossary:domain`. 3

CHEN, I. & FOLOPPE, N. (2011). Is conformational sampling of drug-like molecules a solved problem? *Drug Dev. Res.*, **72**, 85–94. 191

CHRISTIANSON, C.V., MONTAVON, T.J., FESTIN, G.M., COOKE, H.A., SHEN, B. & BRUNER, S.D. (2007a). The mechanism of MIO-based aminomutases in beta-amino acid biosynthesis. *J Am Chem Soc*, **129**, 15744–5. 94, 96

CHRISTIANSON, C.V., MONTAVON, T.J., VAN LANEN, S.G., SHEN, B. & BRUNER, S.D. (2007b). The structure of l-tyrosine 2,3-aminomutase from the c-1027 enediyne antitumor antibiotic biosynthetic pathway. *Biochemistry*, **46**, 7205–7214. 7, 138

CHRISTIANSON, C.V., MONTAVON, T.J., VAN LANEN, S.G., SHEN, B. & BRUNER, S.D. (2007c). The structure of l-tyrosine 2,3-aminomutase from the C-1027 enediyne antitumor antibiotic biosynthetic pathway. *Biochemistry*, **46**, 7205–14. 94

CHRISTIANSON, C.V., MONTAVON, T.J., VAN LANEN, S.G., SHEN, B. & BRUNER, S.D. (2007d). The structure of l-tyrosine 2,3-aminomutase from the C-1027 enediyne antitumor antibiotic biosynthetic pathway. *Biochemistry*, **46**, 7205–14. 97

CIRINO, P.C. & SUN, L. (2008). Advancing biocatalysis through enzyme, cellular, and platform engineering. *Biotechnol Prog*, **24**, 515–9. 21

Cluis, C.P., Burja, A.M. & Martin, V.J. (2007). Current prospects for the production of coenzyme Q10 in microbes. *Trends Biotechnol*, **25**, 514–21. 116

Cochrane, J.C. & Strobel, S.A. (2008). Riboswitch effectors as protein enzyme cofactors. *RNA*, **14**, 993–1002. 96

Coward-Kelly, G. & Chen, R.R. (2008). A window into biocatalysis and biotransformations. *Biotechnol Prog*, **23**, 52–4. 21

Dalton, T.P., Chen, Y., Schneider, S.N., Nebert, D.W. & Shertzer, H.G. (2004). Genetically altered mice to evaluate glutathione homeostasis in health and disease. *Free Radic Biol Med*, **37**, 1511–26. 80, 81, 82

Das, U., Chen, S., Fuxreiter, M., Vaguine, A.A., Richelle, J., Berman, H.M. & Wodak, S.J. (2001). Checking nucleic acid crystal structures. *Acta Crystallographica Section D*, **57**, 813–828. 30

Datta, A. (1970). Regulatory role of adenosine triphosphate on hog kidney n-acetyl-d-glucosamine 2-epimerase. *Biochemistry*, **9**, 3363–3370. 8

Davidson, V. (2004). Electron transfer in quinoproteins. *Arch Biochem Biophys*, **428**, 32–40. 104

Davis, A.M., St-Gallay, S.A. & Kleywegt, G.J. (2008). Limitations and lessons in the use of x-ray structural information in drug design. *Drug Discov Today.*, **13**, 831–41. 151, 153

Daylight Chemical Information Systems (2008). Daylight Theory SMILES. http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html. 53

Debye, P. (1913). Interferenz von rontgenstrahlen und warmebewegung. *Ann. Phys.*, **348**, 49–92. 30

Degtyarenko, K., deMatos, P.D., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. & Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, **36**, D344–350. 24, 33, 36, 56, 59

DEY, M., LI, X., KUNZ, R.C. & RAGSDALE, S.W. (2010). Detection of organometallic and radical intermediates in the catalytic mechanism of methyl-coenzyme M reductase using the natural substrate methyl-coenzyme M and a coenzyme B substrate analogue. *Biochemistry*, **49**, 10902–11. 22

DICKMAN, S.R. (1977). Ribonucleotide reduction and the possible role of cobalamin in evolution. *J. Mol. Evol.*, **10**, 251–60. 86

DUBOIS, J. & KLINMAN, J. (2005). Mechanism of post-translational quinone formation in copper amine oxidases and its relationship to the catalytic turnover. *Arch Biochem Biophys*, **433**, 255–65. 104, 105, 106, 107

DUINE, J., VAN DER MEER, R. & GROEN, B. (1990). The cofactor pyrroloquinoline quinone. *Annu Rev Nutr*, **10**, 297–318. 103, 104

DUINE, J.A. (2001). Cofactor diversity in biological oxidations: implications and applications. *Chemical record (New York, N.Y.)*, **1**, 74–83. 75, 106, 107

DUNATHAN, H.C. (1966). Conformation and reaction specificity in pyridoxal phosphate enzymes. *Proc Natl Acad Sci U S A*, **55**, 712–6. 78

ERMLER, U. (2005). On the mechanism of methyl-coenzyme M reductase. *Dalton Trans*, 3451–8. 98

ESPOSITO, L., VITAGLIANO, L. & MAZZARELLA, L. (2002). Recent advances in atomic resolution protein crystallography. *Protein Pept Lett*, **9**, 95–106. 31

FELTON, L. & ANTHONY, C. (2005). Biochemistry: role of PQQ as a mammalian enzyme cofactor? *Nature*, **433**, E10; discussion E11–2. 104

FERNANDES, P. (2010). Miniaturization in biocatalysis. *Int J Mol Sci*, **11**, 858–79. 21

FERSHT, A. (1999). *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding.*. W. H. Freeman and Company, New York, 1st edn. 2, 3, 20

Fielding, A.H. (2007). *Cluster and classification techniques for the biosciences*. Cambridge University Press. 45, 46, 48, 49, 50, 51

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R. & Bateman, A. (2010). The pfam protein families database. *Nucleic Acids Res*, **38**, D211–22. 3

Fischer, J.D., Holliday, G.L., Rahman, S.A. & Thornton, J.M. (2010a). The structures and physicochemical properties of organic cofactors in biocatalysis. *J Mol Biol*, **403**, 803–24. 112

Fischer, J.D., Holliday, G.L. & Thornton, J.M. (2010b). The CoFactor database: organic cofactors in enzyme catalysis. *Bioinformatics*, **26**, 2496–2497. 41, 56, 146

Fischer, M. & Bacher, A. (2005). Biosynthesis of flavocoenzymes. *Nat Prod Rep*, **22**, 324–50. 73

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K.B., Bairoch, A., Schomburg, D., Tipton, K.F. & Apweiler, R. (2004). Intenz, the integrated relational enzyme database. *J Mol Biol*, **32**, D434–437. 25, 33, 56

Fontaine, F. (2009). OpenBabel: obprop. http://openbabel.org/wiki/Obprop. 42

Frank, R.A., Leeper, F.J. & Luisi, B.F. (2007). Structure, mechanism and catalytic duality of thiamine-dependent enzymes. *Cellular and molecular life sciences : CMLS*, **64**, 892–905. 71, 72

Frank, R.A., Kay, C.W., Hirst, J. & Luisi, B.F. (2008). Off-pathway, oxygen-dependent thiamine radical in the krebs cycle. *J Am Chem Soc*, **130**, 1662–1668. 72

Fries, M., Jung, H.I. & Perham, R.N. (2003). Reaction mechanism of the heterotetrameric (alpha2beta2) E1 component of 2-oxo acid dehydrogenase multienzyme complexes. *Biochemistry*, **42**, 6996–7002. 107

FURUYAMA, K., KANEKO, K. & VARGAS, P. (2007). Heme as a magnificent molecule with multiple missions: heme determines its own fate and governs cellular homeostasis. *Tohoku J Exp Med*, **213**, 1–16. 100, 102

GALPERIN, M.Y., WALKER, D.R. & KOONIN, E.V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res*, **8**, 779–90. 5

GAN, G. (2007). *Data clustering: theory, algorithms, and applications*. ASA-SIAM. 50

GHERARDINI, P.F., WASS, M.N., HELMER-CITTERICH, M. & STERNBERG, M.J. (2007). Convergent evolution of enzyme active sites is not a rare phenomenon. *J Mol Biol*, **372**, 817–45. 36

GHOSE, A.K., VISWANADHAN, V.N. & WENDOLOSKI, J.J. (1998). Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *The Journal of Physical Chemistry A*, **102**, 3762–3772. 43

GOLOVIN, A., DIMITROPOULOS, D., OLDFIELD, T., RACHEDI, A. & HENRICK, K. (2005). MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites. *Proteins*, **58**, 190–9. 24

GRAMATIKOVA, S., MOURATOU, B., STETEFELD, J., MEHTA, P.K. & CHRISTEN, P. (2002). Pyridoxal-5'-phosphate-dependent catalytic antibodies. *J Immunol Methods*, **269**, 99–110. 80

GREENE, L.H., LEWIS, T.E., ADDOU, S., CUFF, A., DALLMAN, T., DIBLEY, M., REDFERN, O., PEARL, F., NAMBUDIRY, R., REID, A., SILLITOE, I., YEATS, C., THORNTON, J.M. & ORENGO, C.A. (2007). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res*, **35**, D291–7. 3, 33, 58

GUILLET, M. & RODRIGUE, N. (2009). Multivariate data analysis school., class Lecture. Creascience Inc. 45, 46, 47

Han, K.F., Bystroff, C. & Baker, D. (1997). Three-dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci*, **6**, 1587–90. 27

Hanson, R.M. (2010). Jmol: a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography*, **43**, 1250–60. 70, 156

Hart, W.E. & Istrail, S. (1997). Robust proofs of NP-hardness for protein folding: general lattices and energy potentials. *J Comput Biol*, **4**, 1–22. 27

Holliday, G.L. (2009). MACiE: Mechanism, Annotation and Classification in Enzymes. http://www.ebi.ac.uk/thornton-srv/databases/MACiE/. 26

Holliday, G.L., Bartlett, G.J., Almonacid, D.E., O'boyle, N.M., Murray-Rust, P., Thornton, J.M. & Mitchell, J.B.O. (2005). Macie: a database of enzyme reaction mechanisms. *Bioinformatics*, **21**, 4315–4316. 8, 26

Holliday, G.L., Almonacid, D.E., Bartlett, G.J., O'Boyle, N.M., Torrance, J.W., Murray-Rust, P., Mitchell, J.B. & Thornton, J.M. (2007a). Macie (mechanism, annotation and classification in enzymes): novel tools for searching catalytic mechanisms. *Nucleic Acids Res*, **35**, D515–20. 26, 32, 33, 40, 58, 61, 142, 146

Holliday, G.L., Thornton, J.M., Marquet, A., Smith, A.G., Rebeille, F., Mendel, R., Schubert, H.L., Lawrence, A.D. & Warren, M.J. (2007b). Evolution of enzymes and pathways for the biosynthesis of cofactors. *Nat Prod Rep*, **24**, 972–87. 77, 78, 85, 86, 87, 90, 91, 96, 140

Holliday, G.L., Mitchell, J.B. & Thornton, J.M. (2009). Understanding the functional roles of amino acid residues in enzyme catalysis. *J Mol Biol*, **390**, 560–77. xvi, 145, 146, 147

Holliday, G.L., Fischer, J.D., Mitchell, J.B.O. & Thornton, J.M. (2011). The complexity of enzymes: from mechanisms to structures. *FEBS J.*, submitted. 5

HOLMES, D., MOODY, P. & DINE, D. (2006). *Research methods for the biosciences*. Oxford University Press. 44

HOLSCHER, T., SCHLEYER, U., MERFORT, M., BRINGER-MEYER, S., GORISCH, H. & SAHM, H. (2009). Glucose oxidation and PQQ-dependent dehydrogenases in gluconobacter oxydans. *J Mol Microbiol Biotechnol*, **16**, 6–13. 104

HUBBARD, S.J. & THORNTON, J.M. (1993). NACCESS. http://www.bioinf.manchester.ac.uk/naccess/, department of Biochemistry and Molecular Biology, University College London. 40

HUSSON, F., JOSSE, J., LE, S. & MAZET, J. (2008). *FactoMineR: Factor Analysis and Data Mining with R*. R package version 1.10. 50

HUTCHISON, G., MORLEY, C. & BANCK, M. (2009). OpenBabel: babel. http://openbabel.org/wiki/Babel. 42

ISHIKAWA, T. & SHIGEOKA, S. (2008). Recent advances in ascorbate biosynthesis and the physiological significance of ascorbate peroxidase in photosynthesizing organisms. *Biosci Biotechnol Biochem*, **72**, 1143–54. 87

IVANISENKO, V.A., PINTUS, S.S., GRIGOROVICH, D.A. & KOLCHANOV, N.A. (2005). PDBSite: a database of the 3D structure of protein functional sites. *Nucleic Acids Res*, **33**, D183–7. 24

JADHAV, V. & YARUS, M. (2002). Coenzymes as coribozymes. *Biochimie*, **84**, 877–88. 23, 150

JOLLIFFE, I.T. (2010). *Principal Component Analysis (Springer Series in Statistics)*. Springer, 2nd edn. 45

JOOSTEN, V. & VANBERKEL, W.J. (2007). Flavoenzymes. *Current opinion in chemical biology*, **11**, 195–202. 73

KABSCH, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Cryst*, **A32**, 922–923. 52

Kahraman, A., Morris, R.J., Laskowski, R.A. & Thornton, J.M. (2007). Shape variation in protein binding pockets and their ligands. *J Mol Biol.*, **368**, 283–301. 4, 152

Kamerlin, S.C., Mavri, J. & Warshel, A. (2010). Examining the case for the effect of barrier compression on tunneling, vibrationally enhanced catalysis, catalytic entropy and related issues. *FEBS Lett*, **584**, 2759–66. 20

Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27–30. 23, 25, 33, 41, 59

Kaplun, A., Binshtein, E., Vyazmensky, M., Steinmetz, A., Barak, Z., Chipman, D.M., Tittmann, K. & Shaanan, B. (2008). Glyoxylate carboligase lacks the canonical active site glutamate of thiamine-dependent enzymes. *Nat Chem Biol*, **4**, 113–8. 71

Kelley, L.A., Gardner, S.P. & Sutcliffe, M.J. (1996). An automated approach for clustering an ensemble of nmr-derived protein structures into conformationally related subfamilies. *Protein engineering*, **9**, 1063–1065. 39, 129

Kim, J., Jia, H. & Wang, P. (2006). Challenges in biocatalysis for enzyme-based biofuel cells. *Biotechnol Adv*, **24**, 296–308. 20

Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wahlby, A. & Jones, T.A. (2004). The uppsala Electron-Density server. *Acta Crystallogr D Biol Crystallogr*, **60**, 2240–9. 30

Klinman, J.P. (2001). How many ways to craft a cofactor? *Proc Natl Acad Sci U S A*, **98**, 14766–8. 97, 105, 106

Klinman, J.P. (2009). An integrated model for enzyme catalysis emerges from studies of hydrogen tunneling. *Chem Phys Lett.*, **471**, 179–93. 20

Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C. & Wishart, D.S. (2011). DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res*, **39**, D1035–41. 25

KOOLMANN, J. & ROHM, K.H. (2003). *Taschenatlas der Biochemie*. Theime Verlag, Stuttgart, 3rd edn. 77

KRISSINEL, E. & HENRICK, K. (2007). Inference of macromolecular assemblies from crystalline state. *J Mol Biol*, **372**, 774–97. 31

LAHAV, N. (2001). The emergence of life on earth. *Progress in Biophysics and Molecular Biology*, **75**, 75–120. 23

LARSSON, A. & REICHARD, P. (1966). Enzymatic synthesis of deoxyribonucleotides. ix. allosteric effects in the reduction of pyrimidine ribonucleotides by the ribonucleoside diphosphate reductase system of escherichia coli. *The Journal of biological chemistry*, **241**, 2533–2539. 5, 8

LASKOWSKI, R.A. (2009). Pdbsum new things. *Nucleic acids research*, **37**, D355–359. 24, 33, 58

LEEPER, F.J. & SMITH, A.G. (2007). Editorial: vitamins and cofactors - chemistry, biochemistry and biology. *Nat Prod Rep*, **24**, 923–6. 72, 83, 84, 85, 86

LIPINSKI, C.A., LOMBARDO, F., DOMINY, B.W. & FEENEY, P.J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*, **46**, 3–26. 113

LITERATURE SERVICES GROUP, EMBL-EBI (2011). Citexplore. `http://www.ebi.ac.uk/citexplore/`. 59, 64

LIU, T., LIN, Y., WEN, X., JORISSEN, R.N. & GILSON, M.K. (2007). BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*, **35**, D198–201. 24

MAECHLER, M., ROUSSEEUW, P., STRUYF, A. & HUBERT, M. (2005). Cluster analysis basics and extensions, rousseeuw et al provided the S original which has been ported to R by Kurt Hornik and has since been enhanced by Martin Maechler: speed improvements, silhouette() functionality, bug fixes, etc. See the 'Changelog' file (in the package source). 51

MANLY, B.F.J. (2004). *Multivariate Statistical Methods: A Primer*. Chapman and Hall/CRC, 3rd edn. 45

MARKUS, R.A. & SUTIN, N. (1985). Electron transfer in chemistry and biology. *Biochim Biophys Acta*, **811**, 265–322. 20

MARQUET, A., BUI, B.T. & FLORENTIN, D. (2001). Biosynthesis of biotin and lipoic acid. *Vitam Horm*, **61**, 51–101. 108, 138

MASCIOCCHI, J., FRAU, G., FANTON, M., STURLESE, M., FLORIS, M., PIREDDU, L., PALLA, P., CEDRATI, F., RODRIGUEZ-TOMÉ, P. & MORO, S. (2009). MMsINC: a large-scale chemoinformatics database. *Nucleic Acids Res*, **37**, D284–90. 25

MCNAUGHT, A.D. & WILKINSON, A. (1997). *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the ≪Gold Book")*.. WileyBlackwell; 2nd Revised edition edition. 5

MEGANATHAN, R. (2001). Biosynthesis of menaquinone (vitamin K2) and ubiquinone (coenzyme Q): a perspective on enzymatic mechanisms. *Vitam Horm*, **61**, 173–218. 89

MEIJERS, R. & CEDERGREN-ZEPPEZAUER, E. (2009). A variety of electrostatic interactions and adducts can activate nad(p) cofactors for hydride transfer. *Chemico-biological interactions*, **178**, 24–8. 75, 76

MEIJERS, R., MORRIS, R.J., ADOLPH, H.W., MERLI, A., LAMZIN, V.S. & CEDERGREN-ZEPPEZAUER, E.S. (2001). On the enzymatic activation of NADH. *The Journal of biological chemistry*, **276**, 9316–9321. 75, 76

MENDEL, R.R., SMITH, A.G., MARQUET, A. & WARREN, M.J. (2007). Metal and cofactor insertion. *Nat Prod Rep*, **24**, 963–71. 85, 87, 90, 91, 92, 100, 102

MINTEER, S.D., LIAW, B.Y. & COONEY, M.J. (2007). Enzyme-based biofuel cells. *Curr Opin Biotechnol*, **18**, 228–34. 22

Muller, I., Kahnert, A., Pape, T., Sheldrick, G.M., Meyer-Klaucke, W., Dierks, T., Kertesz, M. & Uson, I. (2004). Crystal structure of the alkylsulfatase AtsK: insights into the catalytic mechanism of the fe(ii) alpha-ketoglutarate-dependent dioxygenase superfamily. *Biochemistry*, **43**, 3075–88. 88

Muller, U.F. (2006). Re-creating an rna world. *Cellular and molecular life sciences : CMLS*, **63**, 1278–93. 22, 138

Mure, M. (2004). Tyrosine-derived quinone cofactors. *Acc. Chem. Res.*, **37**, 131–9. 105, 106

Nagano, N. (2005). EzCatDB: the enzyme catalytic-mechanism database. *Nucleic acids research*, **33**, D407–412. 26

Nair, N. & Zhao, H. (2007). Biochemical characterization of an L-Xylulose reductase from neurospora crassa. *Appl Environ Microbiol*, **73**, 2001–4. 22

NC-IUBMB (2011). Enzyme nomenclature. http://www.chem.qmul.ac.uk/iubmb/enzyme/. 25

NC-IUBMB & Webb, E.C. (1992). *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992*. Academic Press. 1, 4, 23, 25, 58, 141

Nordlund, P. & Reichard, P. (2006). Ribonucleotide reductases. *Annual Review of Biochemistry*, **75**, 681–706. 86

Oldenburg, J., Marinova, M., Muller-Reible, C. & Watzka, M. (2008). The vitamin K cycle. *Vitam Horm*, **78**, 35–62. 88, 89

OpenEye Scientific Software (2004–2010). OMEGA version 2.4.1, built 20100616. http://www.eyesopen.com/omega. 53, 154, 193

OpenEye Scientific Software (2011). FAQ. http://demo.eyesopen.com/support/faq/. 53

Overington, J.P., Al-Lazikani, B. & Hopkins, A.L. (2006). How many drug targets are there? *Nat Rev Drug Discov*, **5**, 993–6. 2

Patient, S., Wieser, D., Kleen, M., Kretschmann, E., Jesus Martin, M. & Apweiler, R. (2008). Uniprotjapi: a remote api for accessing uniprot data. *Bioinformatics*, **24**, 1321–1322. 36, 65

Patthy, L. (1999). *Protein evolution.*. Blackwell science Ltd, Oxford, 1st edn. 2, 3

Pearson, A.R., De La Mora-Rey, T., Graichen, M.E., Wang, Y., Jones, L.H., Marimanikkupam, S., Agger, S.A., Grimsrud, P.A., Davidson, V.L. & Wilmot, C.M. (2004). Further insights into quinone cofactor biogenesis: probing the role of maug in methylamine dehydrogenase tryptophan tryptophylquinone formation. *Biochemistry*, **43**, 5494–502. 107

Pegg, S.C., Brown, S.D., Ojha, S., Seffernick, J., Meng, E.C., Morris, J.H., Chang, P.J., Huang, C.C., Ferrin, T.E. & Babbitt, P.C. (2006). Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. *Biochemistry*, **45**, 2545–2555. 26

Percudani, R. & Peracchi, A. (2003). A genomic overview of pyridoxal-phosphate-dependent enzymes. *EMBO Rep*, **4**, 850–4. 78, 80

Plaga, W., Vielhaber, G., Wallach, J. & Knappe, J. (2000). Modification of cys-418 of pyruvate formate-lyase by methacrylic acid, based on its radical mechanism. *FEBS Lett*, **466**, 45–8. 138

Pompella, A., Visvikis, A., Paolicchi, A., De Tata, V. & Casini, A.F. (2003). The changing faces of glutathione, a cellular protagonist. *Biochem Pharmacol*, **66**, 1499–503. 81

Porter, C.T., Bartlett, G.J. & Thornton, J.M. (2004). The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*, **32**, D129–133. 25

Prabhakar, R. & Siegbahn, P. (2004). A theoretical study of the mechanism for the biogenesis of cofactor topaquinone in copper amine oxidases. *J Am Chem Soc*, **126**, 3996–4006. 105

Quinn, G.P. & Keough, M.J. (2002). *Experimental design and data analysis for biologists*. Cambridge University Press. 45

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. 44

Rahman, S.A., Bashton, M., Holliday, G.L., Schrader, R. & Thornton, J.M. (2009). Small molecule subgraph detector (smsd) toolkit. *J Cheminform*, **1**. 44, 54, 133

RCSB Protein Data Bank (2011). Glossary of technical terms. http://www.rcsb.org/pdb/static.do?p=help/glossary.html. 30

Rebeille, F., Ravanel, S., Marquet, A., Mendel, R.R., Webb, M.E., Smith, A.G. & Warren, M.J. (2007). Roles of vitamins B5, B8, B9, B12 and molybdenum cofactor at cellular and organismal levels. *Nat Prod Rep*, **24**, 949–62. 77, 78, 83, 90, 91, 92, 96

Reed, L.J. (1974). Multienzyme complexes. *Accounts of Chemical Research*, **7**, 40–46. 107

Reitz, M., Sacher, O., Tarkhov, A., Trumbach, D. & Gasteiger, J. (2004). Enabling the exploration of biochemical pathways. *Organic & biomolecular chemistry*, **2**, 3226–3237. 26

Rhodes, G. (2006). *Crystallography made crystal clear–a guide for users of macromolecular models*. Academic Press Publications London, 3rd edn. 28, 29

Rizzi, M. & Schindelin, H. (2002). Structural biology of enzymes involved in NAD and molybdenum cofactor biosynthesis. *Current opinion in structural biology*, **12**, 709–720. 76, 91

Roje, S. (2006). S-Adenosyl-L-methionine: beyond the universal methyl group donor. *Phytochemistry*, **67**, 1686–98. 96, 97

Rubin-Pitel, S.B. & Zhao, H. (2006). Recent advances in biocatalysis by directed enzyme evolution. *Comb Chem High Throughput Screen*, **9**, 247–57. 21

Saranya, N. & Selvaraj, S. (2009). Variation of protein binding cavity volume and ligand volume in protein-ligand complexes. *Bioorg Med Chem Lett*, **19**, 5769–72. 152

Sayle, R.A. & Milner-White, E.J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem Sci*, **20**, 374. 70

Schmidt, T.S. & Alp, N.J. (2007). Mechanisms for the role of tetrahydrobiopterin in endothelial function and vascular disease. *Clin Sci (Lond)*, **113**, 47–63. 93

Schomburg, I., Chang, A. & Schomburg, D. (2002). Brenda, enzyme data and metabolic information. *Nucleic Acids Res*, **30**, 47–49. 25

Schwartz, B. & Klinman, J.P. (2001). Mechanisms of biosynthesis of protein-derived redox cofactors. *Vitamins and hormones*, **61**, 219–239. 7

Scott, D.E., Ciulli, A. & Abell, C. (2007). Coenzyme biosynthesis: enzyme mechanism, structure and inhibition. *Natural product reports*, **24**, 1009–1026. 73, 80, 83

Shaanan, B. & Chipman, D.M. (2009). Reaction mechanisms of thiamin diphosphate enzymes: new insights into the role of a conserved glutamate residue. *FEBS J*, **276**, 2447–53. 71

Shin, J.M. & Cho, D.H. (2005). PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res*, **33**, D238–41. 24

Shoolingin-Jordan, P.M. (1995). Porphobilinogen deaminase and uroporphyrinogen III synthase: structure, molecular biology, and mechanism. *J Bioenerg Biomembr*, **27**, 181–95. 103

Sieber, S.A. & Marahiel, M.A. (2005). Molecular mechanisms underlying nonribosomal peptide synthesis: approaches to new antibiotics. *Chem Rev*, **105**, 715–38. 21, 77

Silverman, R.B. (2004). *The organic chemistry of drug design and drug action.*. Elsevier academic press, San Diego, 2nd edn. 4

Silverman, R.B. & Nandi, D.L. (1988). Reduced thioredoxin: a possible physiological cofactor for vitamin k epoxide reductase. further support for an active site disulfide. *Biochemical and biophysical research communications*, **155**, 1248–1254. 5

Simmons, C.R., Magee, C.L., Smith, D.A., Lauman, L., Chaput, J.C. & Allen, J.P. (2010). Three-Dimensional structures reveal multiple ADP/ATP binding modes for a synthetic class of artificial proteins,. *Biochemistry*, **49**, 8689–8699. 152

Sippel, K.H., Robbins, A.H., Reutzel, R., Domsic, J., Boehlein, S.K., Govindasamy, L., Agbandje-McKenna, M., Rosser, C.J. & McKenna, R. (2008). Structure determination of the cancer-associated mycoplasma hyorhinis protein mh-p37. *Acta Crystallogr D Biol Crystallogr*, **64**, 1172–8. xvii, 163, 164

Smirnoff, N. (2000). Ascorbate biosynthesis and function in photoprotection. *Philos Trans R Soc Lond B Biol Sci*, **355**, 1455–64. 87, 88

Smith, L.J., Kahraman, A. & Thornton, J.M. (2010). Heme proteins–diversity in structural characteristics, function, and folding. *Proteins*, **78**, 2349–68. xiii, 100, 101

Soda, K., Yoshimura, T. & Esaki, N. (2001). Stereospecificity for the hydrogen transfer of pyridoxal enzyme reactions. *Chem Rec*, **1**, 373–84. 78, 79

Soding, J., Biegert, A. & Lupas, A. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, **33**, W244–W248. 27

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. (2003). The chemistry development kit (cdk): an open-source java library for chemo- and bioinformatics. *J Chem Inf Comput Sci*, **43**, 493–500. 42

Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R. & Willighagen, E.L. (2006). Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr Pharm Des*, **12**, 2111–20. 42

Stockwell, G.R. & Thornton, J.M. (2006). Conformational diversity of ligands bound to proteins. *J Mol Biol.*, **356**, 928–944. 152, 153, 174, 193, 194

Stout, G.H. & Jensen, L.H. (1989). *X-ray structure determination: a practical guide*. Wiley-Blackwell, 2nd edn. 29

Sun, D., Ono, K., Okajima, T., Tanizawa, K., Uchida, M., Yamamoto, Y., Mathews, F.S. & Davidson, V.L. (2003). Chemical and kinetic reaction mechanisms of quinohemoprotein amine dehydrogenase from paracoccus denitrificans. *Biochemistry*, **42**, 10896–903. 107

Sutcliffe, M.J. & Scrutton, N.S. (2000). Enzyme catalysis: over-the-barrier or through-the-barrier? *Trends Biochem Sci.*, **25**, 405–8. 20

Taga, M.E. & Walker, G.C. (2008). Pseudo-B12 joins the cofactor family. *J Bacteriol*, **190**, 1157–9. 85

Thony, B., Auerbach, G. & Blau, N. (2000). Tetrahydrobiopterin biosynthesis, regeneration and functions. *Biochem J*, **347 Pt 1**, 1–16. 92, 93, 94

Tittmann, K., Mesch, K., Pohl, M. & Hubner, G. (1998). Activation of thiamine diphosphate in pyruvate decarboxylase from zymomonas mobilis. *FEBS Lett*, **441**, 404–6. 140

Tran, U.C. & Clarke, C.F. (2007). Endogenous synthesis of coenzyme Q in eukaryotes. *Mitochondrion*, **7 Suppl**, S62–71. 89

Tripathy, B.C., Sherameti, I. & Oelmuller, R. (2010). Siroheme: an essential component for life on earth. *Plant Signal Behav*, **5**, 14–20. 100, 102

Tsai, T.Y., Yang, C.Y., Shih, H.L., Wang, A.H. & Chou, S.H. (2009). Xanthomonas campestris PqqD in the pyrroloquinoline quinone biosynthesis operon adopts a novel saddle-like fold that possibly serves as a PQQ carrier. *Proteins*, **76**, 1042–8. 106

Turner, M.A., Yang, X., Yin, D., Kuczera, K., Borchardt, R.T. & Howell, P.L. (2000). Structure and function of s-adenosylhomocysteine hydrolase. *Cell Biochem Biophys*, **33**, 101–25. 97

Union of Pure, I. & Applied Chemistry (2005–2009). IUPAC compendium of chemical terminology - the gold book. http://goldbook.iupac.org/. 5, 6, 7

UniProt Consortium (2010). The universal protein resource (UniProt) in 2010. *Nucleic Acids Res*, **38**, D142–8. 33

Unno, M., Matsui, T. & Ikeda-Saito, M. (2007). Structure and catalytic mechanism of heme oxygenase. *Nat Prod Rep*, **24**, 553–70. 100, 102

Vainio, M.J. & Johnson, M.S. (2007). Generating conformer ensembles using a multiobjective genetic algorithm. *Journal of Chemical Information and Modeling*, **47**, 2462–74. 152

van der Palen, C.J., Slotboom, D.J., Jongejan, L., Reijnders, W.N., Harms, N., Duine, J.A. & van Spanning, R.J. (1995). Mutational analysis of mau genes involved in methylamine metabolism in paracoccus denitrificans. *Eur J Biochem*, **230**, 860–71. 107

Van Petegem, F., De Vos, D., Savvides, S., Vergauwen, B. & Van Beeumen, J. (2007). Understanding nicotinamide dinucleotide cofactor and substrate specificity in class I flavoprotein disulfide oxidoreductases:

crystallographic analysis of a glutathione amide reductase. *J Mol Biol*, **374**, 883–9. 80, 82

Varoquaux, G. (2011). Data processing – identifying relevant variables: Pca and ica. http://gael-varoquaux.info/scientific_computing/ica_pca/index.html. 45

Wallace, A.C., Laskowski, R.A. & Thornton, J.M. (1995). LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*, **8**, 127–34. 65

Waller, I. (1923). Zur frage der einwirkung der warmebewegung auf die interferenz von rontgenstrahlen. *Zeitschrift fur Physik A - Hadrons and Nuclei*, **17**, 398–408. 30

Wang, R., Fang, X., Lu, Y., Yang, C.Y. & Wang, S. (2005). The PDBbind database: methodologies and updates. *J Med Chem*, **48**, 4111–9. 24

Ward Jr, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 236–244. 129

Warren, M.J. & Scott, A.I. (1990). Tetrapyrrole assembly and modification into the ligands of biologically functional cofactors. *Trends Biochem Sci*, **15**, 486–91. 103

Webb, M.E., Marquet, A., Mendel, R.R., Rebeille, F. & Smith, A.G. (2007). Elucidating biosynthetic pathways for vitamins and cofactors. *Nat Prod Rep*, **24**, 988–1008. 80, 83

Wei, C.C., Crane, B.R. & Stuehr, D.J. (2003). Tetrahydrobiopterin radical enzymology. *Chem Rev*, **103**, 2365–83. 91, 92, 93

Weng, Y.Z., Chang, D., Huang, Y.F. & Lin, C.W. (2011). A study on the flexibility of enzyme active sites. *BMC Bioinformatics*, **12**, S32. 152

WILLETT, P., BARNARD, J.M. & DOWNS, G.M. (1998). Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, **38**, 983–996. 51, 52

WOGGON, W.D., WAGENKNECHT, H.A. & CLAUDE, C. (2001). Synthetic active site analogues of heme-thiolate proteins. characterization and identification of intermediates of the catalytic cycles of cytochrome p450cam and chloroperoxidase. *J Inorg Biochem*, **83**, 289–300. 102

YAMAGUCHI, A., IIDA, K., MATSUI, N., TOMODA, S., YURA, K. & GO, M. (2004). Het-PDB navi.: a database for protein-small molecule interactions. *J Biochem*, **135**, 79–84. 24

YOSHIDA, K., YAMAGUCHI, M., MORINAGA, T., KINEHARA, M., IKEUCHI, M., ASHIDA, H. & FUJITA, Y. (2008). Myo-inositol catabolism in bacillus subtilis. *J Biol Chem*, **283**, 10415–24. 72

ZHOU, H., XIE, X. & TANG, Y. (2008). Engineering natural products using combinatorial biosynthesis and biocatalysis. *Curr Opin Biotechnol*, **19**, 590–6. 21

ZHU, Y. & SILVERMAN, R.B. (2008). Revisiting heme mechanisms. a perspective on the mechanisms of nitric oxide synthase (NOS), heme oxygenase (HO), and cytochrome P450s (CYP450s). *Biochemistry*, **47**, 2231–43. 98, 100