Evolution study of the Baeyer-Villiger monooxygenases enzyme family: functional importance of the highly conserved residues.

Joseph Rebehmed, Véronique Alphand, Véronique de Berardinis, Alexandre de Brevern

Evolution study of the Baeyer-Villiger monooxygenases enzyme family: functional importance of the highly conserved residues.

Joseph REBEHMED[1,2,3,4], Véronique ALPHAND[5] , Véronique DE BERARDINIS[6] and Alexandre G. DE BREVERN[1,2,3,4]*

[1] INSERM, U665, DSIMB, F-75739 Paris, France

[2] Univ. Paris Diderot, Sorbonne Paris Cité, UMR_S 665, F-75739 Paris, France

[3] Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France

[4] Laboratoire d'Excellence GR-Ex, F-75739 Paris, France

[5] Institut des Sciences Moléculaires de Marseille, UMR CNRS 7313, Université Aix-Marseille, Marseille, France

[6] CEA, DSV, IG, Genoscope, 2 rue Gaston Crémieux, 91057 Evry, France

* Corresponding authors:

Mailing address: de Brevern Alexandre G., INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Denis Diderot, Sorbonne Paris Cité, INTS, 6 rue Alexandre Cabanel, 75739 Paris Cedex 15, France.

E-mail: alexandre.debrevern@univ-paris-diderot.fr

Tel: +33(1) 44 49 30 38 - Fax: +33(1) 47 34 74 31

Mailing address: Rebehmed Joseph., INSERM UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Denis Diderot, Sorbonne Paris Cité, INTS, 6 rue Alexandre Cabanel, 75739 Paris Cedex 15, France.

E-mail: joseph.rebehmed@univ-paris-diderot.fr

Tel: +33(1) 44 49 30 91 - Fax: +33(1) 47 34 74 31

Abstract

Baeyer-Villiger monooxygenases (BVMOs) catalyze the transformation of linear and cyclic ketones into their corresponding esters and lactones by introducing an oxygen atom into a C–C bond. This bioreaction has numerous advantages compared to its chemical version; it does not induce the use of potentially harmful reagents (*i.e.*, green chemistry) and displays significant better enantio- and regio- selectivity.

New potential BVMOs were searched using sequence homology for type I BVMO proteins. 116 new sequences were identified as new putative BVMOs respecting the defined selection criteria. Multiple sequence alignments were carried out on the selected sequences to study the conservation of structurally and/or functionally important amino acids during evolution. Type I BVMO signature motif was found to be conserved in 94.8% of the sequences. We noticed also the highly conserved – but previously unnoticed – Threonine 167 (93.1%), located in the signature motif; this position could be added in the pattern used to characterize specific Type I enzymes. Amino acids at the vicinity of the FAD and NADPH cofactors were found also to be highly conserved and the details of the interactions were emphasized. Interestingly, residues at the enzyme binding site were found less conserved in terms of sequence evolution, leading sometimes to some important amino acid changes. These behaviors could explain the enzyme selectivity and specificity for different ligands.

Key-words: sequence alignment; phylogeny; protein structure; function; interactions.

1 Introduction

Baeyer-Villiger (BV) oxidation is a useful synthetic tool in organic chemistry. It transforms linear and cyclic ketones into their corresponding esters or lactones by introducing an oxygen atom into a carbon-carbon bond [1]. This reaction can also be performed via biocatalysis by a family of enzymes called Baeyer-Villiger Monooxygenases (BVMOs) [2]. This bioreaction displayed better chemo-, regio- and enantio- selectivity than the chemical version; it also meets the demands of sustainable and green chemistry [3, 4].

BVMOs are flavoenzymes and belong to the class of oxidoreductases [2, 5]. They catalyze the oxidation of linear, cyclic and aromatic ketones to esters or lactones respectively. During enzymatic oxidation, one atom of oxygen is incorporated between a carbon-carbon bond, whereas the other oxygen atom ends up in a water molecule with the hydrogen atoms originating from the cofactor NAD(P)H. BVMOs are typically soluble proteins and work without additional proteins [6, 7].

BVMOs contain a flavin cofactor (FAD or FMN), which is crucial for catalysis and is tightly, but not covalently, bound in the active site. Furthermore, these enzymes require the reduction of the flavin cofactor to activate it for molecular oxygen binding. Either NADH or NADPH acts as the electron donor. In 1997, Willetts observed that at least two classes of BVMOs exist [8]. Type I BVMOs, which are the most intensively studied, consist of only one polypeptide chain and are FAD and NADPH-dependent for their activity. Type II BVMOs use FMN as flavin cofactor and NADH as electron donor and are composed of two different subunits. Most biochemical and biocatalytic studies [9] have been performed on the cyclohexanone monooxygenase (CHMO) from *Acinetobacter calcoaceticus* NCIMB 9871 (E.C. 1.14.13.22) [10] that belongs to type I BVMOs. This CHMO was shown to be active

against a remarkable number of substrates (over 100 different substrates) exhibiting an exquisite chemo-, enantio- and/or regio- selectivity [2, 11].

1.1. Enzymatic mechanism.

The tightly bound FAD molecule is reduced by NADPH. The reduced flavin then reacts with molecular oxygen to form a C4A-peroxyflavin intermediate (E•FADHOO^{-}•NADP^{+}). This peroxyflavin intermediate plays the same role as the peracid in the conventional Baeyer-Villiger oxidation in organic chemistry and will react with a ketone substrate. This produces the tetrahedral Criegee intermediate that subsequently rearranges to give the C4a-hydroxyflavin and the lactone or ester products [6, 12]. A molecule of water is spontaneously eliminated from the hydroxyflavin to regenerate the oxidized FAD. Figure 1 presents a simplified scheme of the catalytic mechanism of Type I BVMOs. For a more detailed and elaborated scheme please see Mirza et al 2009 [13].

1.2. Structural features.

The type I BVMOs consist of only one polypeptide chain. In 2004, Mattevi et al. solved the first crystal structure of a BVMO, PAMO (Phenylacetone monooxygenase) from *Thermofibida fusca* (PDB code 1W4X) [14]. Subsequently, the structure of CHMO from *Rhodococcus* sp. HI-31 was crystallized in complex with both FAD and NADPH by Mirza et al. in 2009 (PDB code 3GWD and 3GWF) [13]. The available structures revealed a two-domain architecture and the active site is located in a cleft at the domain interface. They contain two Rossmann-folds [15] (with sequence motif GxGxxG which is frequently occurring in nucleotide binding proteins) indicating that these enzymes bind two cofactors using separate dinucleotide binding domains: one for the FAD and the other for NADPH cofactor [16]. Figure 2 presents the overall structure of CHMO enzyme (PDB code 3GWD)

[13] with both cofactors shown in sticks and transparent surface. The FAD cofactor is tightly bound and buried within the FAD domain (green), in agreement with the observation that this cofactor does not dissociate from the enzyme, while the NADPH is sandwiched between the Rossmann fold of the NADPH domain (blue).

The main purpose of this work is the search for new potential BVMOs to expand the enzymatic toolbox for biocatalysis and cover a broader range of substrates types and enantio- and regio- selectivity. We will focus on the Type I BVMO subclass and will give an overview of the sequence, structure and/or function relationships. Amino acids conservation for the different type I BVMOs related sequences during evolution will be examined. Residues highly conserved and playing important roles for the structure and function of the enzyme will be highlighted. The details of the interactions between the enzymes and the different cofactors will be emphasized.

2 Materials and Methods

The protocol used to look for BVMO type I related sequence is summarized in the Figure 3. The sequences of three different type I BVMO enzymes were obtained from the Universal Protein Resource (UniProt) [17, 18] website (www.uniprot.org) and used as query: CHMO from *Rhodococcus* sp. HI-31 (UniProt ID: C0STX7), PAMO from *T. fusca* [19] (UniProt ID: Q47PU3) and cyclopentadecanone monooxygenase (CPDMO) from *Pseudomonas* sp. HI-70 [20] (UniProt ID: Q1T7B5). They were chosen because their substrate profiles are different and together cover a large scope of compounds. Two of them have at least one crystal structure deposited in the Protein Data Bank (Figure 3.1). In a first step, PSI-BLAST (version 2.2) [21] was used to look for related sequences in the UniRef90 database (March 2012 release) [22] (Figure 3.2). The potential homologous proteins were selected using defined criteria of E-value and sequence identity with the corresponding query. Unrelated, incomplete

and redundant sequences from the different searches were removed from the final homologues pool before continuing to the next step (Figure 3.3).

Multiple sequence alignments (MSA) were made with ClustalW2 (version 2.1) [23] and Muscle (version 3.8) [24] programs. Sequence alignments were visualized using the Jalview program (version 2.7) [25] (Figure 3.4). PhyML software (version 3.0) [26] was used to estimate maximum likelihood phylogenies from the protein sequence alignments (Figure 3.5). Tree representations of the phylogenic results were generated by Dendroscope software (version 3.2) [27]. The conservation score of each residue was calculated using Rate4Site (version 2.01) [28] (Figure 3.6) and the three-dimensional structure of the enzyme was colored according to the conservation scores using PyMOL (The PyMOL Molecular Graphics System, Version 1.5, Shrödinger, LLC) (Figure 3.7).

3 Results and Discussion

3.1 Search for BVMO related sequences in databases

The UniProt reference clusters (UniRef) combine closely related sequences into a single record to reduce database size and speed up significantly sequence similarity searches. Various non-redundant databases with different sequence identity cut-offs exist. In this work, we used the Uniref90 database, in which no pair of sequences in the representative set has more than 90% mutual sequence identity respectively [29]. The protein sequence of three type I BVMOs, CHMO from *Rhodococcus* sp. HI-31 (UniProt ID: C0STX7; PDB code: 3GWD), PAMO from *T. fusca* [19] (UniProt ID: Q47PU3; PDB code: 1W4X) and cyclopentadecanone monooxygenase (CPDMO) from *Pseudomonas* sp. HI-70 [20] (UniProt ID: Q1T7B5) were used as query with the PSI-BLAST program. We performed three consecutive iterations of similarity searches. Redundant and unrelated sequences were removed from the set. The database search yielded 19, 71 and 34 proteins sequences respectively respecting the defined

criteria (E-value smaller than e^{-10} and sequence identity with the query higher than 50% to ensure related sequences). Eight sequences were as well removed from the set and were not taken into consideration in the analyses because they consist of fragments (incomplete) or their sequence lengths are too long (more than 850 residues). Finally, we obtained 116 sequences between 500 and 653 residues long that will proceed to the multiple sequence alignments.

3.2 Multiple sequence alignments (MSA)

To compare the sequences obtained from the PSI-BLAST search among each other and with their corresponding queries, pairwise and multiple sequence alignments were performed with ClustalW2 and MUSCLE program respectively. MUSCLE is known to accomplish faster and more accurate multiple sequence alignments than ClustalW2 however it does not perform pairwise alignments. The remaining analyses of conservation score calculation and phylogeny will be based on the multiple sequence alignments obtained by MUSCLE program and the average sequence identities will be calculated from the pairwise alignments obtained by ClustalW2.

Conserved residues are usually involved in protein function or structural stability. Residues numbering will be based on the protein sequence of CHMO (540 amino acids long) from *Rhodococcus* sp. HI-31 to avoid any confusion while analyzing and presenting the results unless it is specified otherwise. MSA revealed several conserved residues that enabled us to identify possible important regions: dinucleotide-binding domains for FAD and NADPH. Both domains incorporate GXGXX(G/A) motifs, which are part of Rossmann fold [30, 31]. The first motif is located between residues 15 and 20 with Gly15, Gly17 found both to be conserved in 99.1% of the sequence while position 20 is represented by a Glycine in 97.4% and 2.6% by an Alanine residue. The second motif, which belongs to the NADPH

domain, lies between residues 185 and 190. Position 185 and 187 are monopolized by Glycine, while position 190 is occupied in 62.1% and 35.3% of the cases by a Glycine and Alanine respectively.

The fingerprint sequence of type I BVMOs subclass (FXGXXXHXXXW(P/D)) [32] is also preserved in our multiple sequence alignment. It is critically involved in the catalysis and hence contains several highly conserved amino acids. It is located between residue 160 and residue 171. Table 1 shows the percentage for the different amino acids composing this fingerprint in our dataset. Phe160 and His166 were found in 94.8% of the sequences while Gly162 and Trp170 are found in all sequences. Position 171 is occupied by a Pro and Asp in 70.7% and 28.4% of the cases respectively. The sequences containing the exact full match of the fingerprint were also counted and the number is equal to 103 over the total of 116 sequences in the dataset. This result shows that the latter is composed of 88.8% type I BVMOs enzymes. Concerning the 13 sequences, only a single residue at the pattern was changed at a time. In six cases, it involves the first residue of the motif (Phenylalanine) that was replaced by another aromatic amino acid (Tyrosine). We also noticed that position 167, which is located inside that highly conserved type I BVMO motif, is a Threonine amino acid in 93.1% of the sequences. Therefore, it is possible to propose a modified type I BVMO signature that includes Thr167: FXGXXXH**T**XXW(P/D). The exact match of the latter motif was found in 82.8% of the sequences (96 over 116). Very recently, Riebel et al. reported in 2012 the presence of a remarkably conserved region located between the N-terminal Rossmann motif and the BVMO motif. This region can be defined by [A/G]GxWxxxx[F/Y]P[G/M]xxxD [33]. The exact match of this pattern was found in 115 sequences in our dataset and is located between position 45 and position 59 for *Rhodococcus* CHMO

3.3 Phylogenetic tree construction

Figure 4 shows the relationship between the sequences as circular cladogram presentation of the maximum likelihood tree generated by PhyML from the multiple sequence alignment obtained by MUSCLE. Three major branches, shown in different colors, can be distinguished; they represent 23.3% for the red cluster that encompasses the CHMO and 29.3% and 47.4% for the green and blue clusters that enclose PAMO and CPDMO respectively. They can be subdivided into clusters of various sizes.

The average pairwise sequence identity of all sequences in our dataset is equal to 40%. It ranges between 49.5% and 54.4% for each cluster underlining some diversity within clusters. Average and range of sequence identity between these different sequence families were also calculated and the results are summarized in Sup. Table 1. Hence, the sequences in the blue branch are evolutionarily more distant from the red and green branches as they share an average sequence identity of 26.4% and 27.4% respectively. The red and green branches share an average sequence identity equal to 44.1%. The radial phylogram located on the top right of Figure 4 also underlines this.

3.4 Conservation score calculation during evolution

Rate4Site is an algorithmic tool for the identification of functionally important regions in proteins by estimating the rate of amino acid substitutions at each position in a MSA, taking into account the evolutionary relations between the homologous proteins, using the maximum likelihood paradigm. Rate4Site uses the MSA previously generated by MUSCLE and the tree obtained from PhyML to detect these important sites. This research is based on the underlying assumptions that, in general, structurally and functionally important residues are slowly evolving. Functionally important residues, *e.g.* in ligand binding and protein-protein interactions, are often evolutionarily conserved and are most likely to be solvent-accessible,

whereas conserved residues within the protein core most probably have an important structural role in maintaining the protein's fold [34, 35]. Thus, estimated evolutionary rates, as well as relative solvent accessibility predictions, are assigned to each amino acid in the sequence; both are subsequently used to indicate residues that have potential structural or functional importance.

From the multiple sequence alignment obtained by MUSCLE, we computed the residue variety in % for each position of *Rhodococcus* CHMO sequence. Similarities between amino acids were also taken into consideration in the analyses. Twenty two positions (4.1%) were found entirely conserved for all the sequences in the dataset, while 47 (8.7%) and 35 (6.5%) of the positions were occupied in 90% to 100% and 80% to 90% of the cases by the same residue respectively. Over 100 positions of 540 have been found preserved in at least 80% of the sequences underlying that Type I BVMO are a highly conserved family of enzymes. Table 1 **s**ummarizes important positions in the type I BVMOs family and the corresponding nature and percentage of residues. The positions that are conserved in more than 70% of the sequences correspond to 21.1% of the residues (see Sup. Table 2).

To elucidate the importance of these conserved residues for the enzyme function and structure, conservation scores will be analyzed on the 3D structure of the enzyme and the existing interactions with both cofactors will be highlighted.

3.5 Projection of the scores on the protein three-dimensional structure

Figure 5 displays *Rhodococcus* CHMO (PDB code 3GWD) structure presented in surface in two different views. The protein was colored according to the conservation score of each residue. The scores are normalized, so that the average score for all residues is zero and the standard deviation is one. The lowest score represents the most conserved position in the protein. The color varies from red for highly conserved residues (normalized conservation

score equal to -1.2) to blue for poorly conserved residues (normalized conservation score equal to 4.0). Moderately conserved residues are colored in white. This presentation will give a better global overview taking into consideration the three-dimensional fold of the protein in the space. For example, the FAD-binding domain of CHMO [13] is composed of residues 1 to 140 and 387 to 540 (green domain in Figure 2), while the NADPH domain is composed of residues 152 to 208 and 335 to 380 (blue domain in Figure 2). In addition, domain detection algorithms were applied on *Rhodococcus* CHMO and also *Thermofibida* PAMO structures using Protein Peeling [36-38] and DIAL [39] webservers. These two enzymes have a Root Mean Square Deviation (RMSD) of 1.4 Å and the obtained results from both approaches showed that Type I BVMO proteins present a similar topology and an organization in two sequentially un-consecutive domains (data from Protein Peeling and DIAL not shown). Therefore, these two domains are not consecutive in the 1D linear amino acid sequence of the enzyme, and only a projection of the scores on the 3D structure will allow visualizing the conserved patches. We can observed that the outer parts of the protein (colored in blue and light blue on the Figure 5) are less conserved than the inner parts which mainly correspond to the FAD and NADPH binding domain as well as the active site that is found in a cleft at the interface of the previous domains. The helical domain (residues 224 to 322 and colored in cyan in Figure 2) does not seem to be highly conserved in the type I BVMO subclass; only few positions were found to be conserved for more than 75% of the sequences. We noticed also that NADPH domain is slightly less conserved than FAD domain.

Figure 6 presents a magnified view of the enzyme active site and cofactor binding domains. This view underlines that all residues interacting with the cofactors are conserved (red color on the figure). The identification and measurements of pockets and cavities in CHMO protein were made with CASTp [40] server. The largest cavity found has a molecular

surface equal to 2306 $Å^2$ and a volume of 3217 $Å^3$. This cavity includes the binding domains of both cofactors which show that they must communicate with each other for the enzymatic activity.

3.6 Highlighting existing interactions

More analyses on an atomistic level were also made to reveal the structural and/or functional role of the conserved residues. PoseView [41, 42] webserver was used on the X-ray structure of CHMO to determine the existing interactions between the enzyme and both cofactors (Figure 7). Important positions and their corresponding residue type and percentages are already summarized in Table 1.

Atom labeling of both cofactors FAD and NADPH is shown in Sup. Figure 1. Reference will be made throughout the coming paragraphs to ease the explanation of the interactions between the enzyme residues and specific atoms of the cofactors. Amino acid at position 19 can be mainly a Glycine or Alanine. This residue makes hydrogen bonds by its backbone amino group with one oxygen atom of the first phosphate group (labeled O2P) of FAD cofactor. Side chain of residues at this position must be small to avoid steric hindrance with FAD. Position 39 is occupied by an Aspartic and Glutamic acid in 20% and 80% respectively. The carboxylic group COO^- of these negatively charged amino acids performs hydrogen bonds with both hydroxyl groups carried by the ribose ring of FAD (O2B and O3B).

The site 47 is filled in 91.3% of the data by a Threonine that creates hydrogen bonds by its backbone amino group with an oxygen of PA phosphate group. An extra interaction can be made through its hydroxyl group side chain with O2' of FAD cofactor. Hydrophobic contacts between this residue and the cofactor are represented more indirectly by means of green contours. More hydrophobic contacts are conducted by Trp48 that is conserved in 100.0% of the sequences.

A Tryptophan residue fills the position 50 in 80.9% of the sequences. As for the Asp39, Trp50 shares through the amino group of its indole ring two hydrogen bonds with both hydroxyl functions of the FAD ribose ring. Otherway it is substituted by other aromatic residues, namely Phenylaline, Tyrosine and Histidine. Position 59 was found to be highly conserved and presented by an Aspartic acid in all the sequences. It makes hydrogen bond through its backbone NH group with the O4 ketone of isoalloxazine ring. The other ketone of the latter ring is engaged in hydrogen bonding with two different residues: Asn436 highly conserved (98.3%) through its side chain amino group and residue at position 437 that is completely variable because this hydrogen bond is made through its backbone amino group. Position 436 was found also to be occupied in 0.9% of the cases by a Glutamine residue which has similar side chain than Asparagine.

Residue 65 is presented in 100% of the cases by a Tyrosine amino acid. The hydroxyl group of its side chain plays the role of hydrogen donor and acceptor and interacts with the hydroxyl groups O3' and O4' of FAD cofactor. Residue at position 112 was found to be occupied by 82.8% of the cases by a Valine and for the remaining sequences by mainly other amino acids with hydrophobic side chains (Alanine, Isoleucine and Methionine). It makes two hydrogen bonds through its backbone with the adenine base of FAD cofactor. The CO and NH groups play the role of acceptor and donor and interact with N6A and N1A atoms respectively.

We analyzed next the residues of the enzyme interacting with the NADPH cofactor. Position 189 can be a Threonine (as in CHMO) or a Serine with the respective percentage of 50.9% and 49.1%. This residue makes a hydrogen bond, through the hydroxyl group of its side chain that acts as a donor, with the PN phosphate group of the NADPH cofactor. One extra hydrogen bond can be made by the amine group of its backbone. Same for position 186 that is occupied in 94.8% of the sequences by a Threonine that acts, through its NH group, as

a hydrogen donor to O3B atom. This position can be also occupied by residues that possess hydrogen bond donor groups (Serine and Asparagine)

The Arg209 is found to be conserved during evolution in more than 100% of the sequences. The analyses of the existing interactions between the protein and the cofactors showed that Arg209 side chain establishes π-cation interaction with the NADPH adenine base as well as electrostatic interactions with the phosphate group of the nucleotide. Position 210 is occupied in 81.9% and 15.5% of the sequences by a Threonine or a Serine respectively. Both amino acids can make hydrogen bond, through their side chain hydroxyl group, with the phosphate (P2B) carried by the ribose ring. Position 492 is occupied by an aromatic amino acid with the possibility of hydrogen bonding by its side chain with the O3D atom of NADPH, i.e. 69.8% of Tryptophane and 28.4% of Tyrosine.

Mirza and co-workers reported the X-ray structure of a *Rhodococcus* CHMO showing an open and a closed form [13]. These conformations revealed domain shifts around multiple linkers and loop movements involving the conserved Arg329 and Trp492. These movements are suggested to be coordinated by the previously mentioned BVMO motif, providing an explanation for the conservation of this sequence motif. Previously, the X-ray structure analysis of PAMO (in the absence of $NADP^+$), reported in 2004 by Malito and co-workers, suggested that Arg337 stabilizes the Criegee intermediate by H-bonding [14].

Comparisons of FAD binding site of *Rhodococcus* CHMO and *T. fusca* PAMO show interesting specificities (see Figure 8). Table 3 lists the corresponding residues in both enzymes. We can observe that both enzyme present similar FAD-binding domain shape and volume. From a total of 15 positions, 9 were found to be the same (60%) in both enzymes. From the residues that differ, some of them are similar and their side chains present the same physico-chemical properties like Asp39 in CHMO that was replaced by Glu46 in PAMO and

therefore they make similar interactions with the FAD cofactor. While for others, the interactions with the cofactor are done primarily through the backbone. Leu437 and Gly19 in CHMO were changed to Met446 and Ser27 respectively in PAMO. Both residues make hydrogen bonding with the FAD cofactor via the amino group of their backbone. And last; the small diversity in the amino acid composition must play an important role in the specificity and selectivity for substrates.

4. Summary and Conclusion

In this work, we looked for related sequences in the UniRef90 database with PSI-BLAST using three well defined type I BVMOs protein sequences query. More than 100 new sequences were identified as putative BVMOs respecting our defined criteria of an E-value smaller than e^{-10} and a sequence identity with the query higher than 50%. These strict selection criteria were chosen as the flavopotein monooxygenase superfamily is suffering from many annotation problems in the databases. Some proteins were cloned and sequenced, but their activities have not been properly tested. Therefore, their submitted names remain vague and unspecific (*i.e.* putative monooxygenase, oxidoreductase). While others enzymes were intensively studied and their corresponding names reveal valuable information on their activities and the substrate they catalyze (i.e. steroid monooxygenase, phenylacetone monooxygenase, cyclohexanone 1,2-monooxygenase).

Multiple sequence alignments were carried out on the selected sequences to study the conservation of structurally and/or functionally important amino acids during evolution. Type I BVMO signature motif was found to be conserved in 94.8% of the sequences. We noticed also the highly conserved Thr167 (93.1%) that is located in the signature motif, and therefore we suggested that the addition of this position in the pattern can be used to distinguish specific Type I enzymes. Both Rossmann Fold motifs were found to be highly conserved as well as

amino acids at the vicinity of the FAD and NADPH cofactors. Residues at the enzyme binding site were less conserved than the previous ones which can explain the enzyme selectivity and specificity for different ligands.

As a conclusion, the identification of new putative BVMOs and the amino acids conservation scores during evolution study combined with structural information will offer very valuable insights for enzyme design and will obviously facilitate the production of novel biocatalysts. It must be noted that all these analyses were made on static structures of the different enzymes. Some residues were found to be highly conserved but not directly involved in the interaction with the cofactors. Treating the structure as flexible using molecular dynamics simulations will reveal the exact functional and/or structural role of these residues as well as the dynamic movement of one domain relative to another and the different reaction lifetime. *in silico* structural models, obtained by comparative modeling and threading techniques, can also be considered for the sequences that do not have experimental structures yet.

5. Acknowledgements and fundings

This work is supported by a grant from the French National Research Agency (ANR): NaturaDyRe (ANR-2010-CD2I-014-04) to Joseph REBEHMED, Alexandre G. DE BREVERN, Véronique DE BERARDINIS and Véronique ALPHAND, from the Ministry of Research (France), University Paris Diderot, Sorbonne Paris Cité (France), National Institute of Blood Transfusion (INTS, France), National Institute of Health and Medical Research (INSERM, France), and “Investissements d’avenir”, Laboratories of Excellence GR-Ex to Joseph REBEHMED and Alexandre G. DE BREVERN, from CEA (France) to Véronique DE BERARDINIS and from Université d’Aix-Marseille (France) and CNRS (France) to Véronique ALPHAND. The authors were granted access to high performance computing

(HPC) resources at the French National Computing Center CINES under grant no. 2012-c2012076930 funded by the GENCI (Grand Equipement National de Calcul Intensif).

6. References

[1] Baeyer A., Villiger V., Einwirkung des Caro'schen Reagens auf Ketone, Berichte der deutschen chemischen Gesellschaft 32 (1899) 3625–3633.
[2] van Berkel W.J., Kamerbeek N.M., Fraaije M.W., Flavoprotein monooxygenases, a diverse class of oxidative biocatalysts, J Biotechnol 124 (2006) 670-689.
[3] Alphand V., Wohlgemuth R., Applications of Baeyer-Villiger Monooxygenases in Organic Synthesis, Current Organic Chemistry 14 (2010) 1928-1965.
[4] Alphand V., Carrea G., Wohlgemuth R., Furstoss R., Woodley J.M., Towards large-scale synthetic applications of Baeyer-Villiger monooxygenases, Trends Biotechnol 21 (2003) 318-323.
[5] Kamerbeek N.M., Janssen D.B., Berkel W.J.H.v., Fraaije M.W., Baeyer-Villiger Monooxygenases, an Emerging Family of Flavin-Dependent Biocatalysts, Advanced Synthesis & Catalysis 345 (2003) 667-678.
[6] Rehdorf J., Mihovilovic M.D., Bornscheuer U.T., Exploiting the regioselectivity of Baeyer-Villiger monooxygenases for the formation of beta-amino acids and beta-amino alcohols, Angew Chem Int Ed Engl 49 (2010) 4506-4508.
[7] Rehdorf J., Mihovilovic M.D., Fraaije M.W., Bornscheuer U.T., Enzymatic synthesis of enantiomerically pure beta-amino ketones, beta-amino esters, and beta-amino alcohols with Baeyer-Villiger monooxygenases, Chemistry 16 (2010) 9525-9535.
[8] Willetts A., Structural studies and synthetic applications of Baeyer-Villiger monooxygenases, Trends Biotechnol 15 (1997) 55-62.
[9] Hilker I., Gutierrez M.C., Furstoss R., Ward J., Wohlgemuth R., Alphand V., Preparative scale Baeyer-Villiger biooxidation at high concentration using recombinant Escherichia coli and in situ substrate feeding and product removal process, Nat Protoc 3 (2008) 546-554.
[10] Chen Y.C., Peoples O.P., Walsh C.T., Acinetobacter cyclohexanone monooxygenase: gene cloning and sequence determination, J Bacteriol 170 (1988) 781-789.
[11] Mihovilovic M.D., Müller B., Stanetty P., Monooxygenase Mediated Baeyer-Villiger Oxidations Eur. J. Org. Chem. (2002) 3711-3730.
[12] Berezina N., Kozma E., Furstoss R., Alphand V., Asymmetric Baeyer-Villiger Biooxidation of a-Substituted Cyanocyclohexanones: Influence of the Substituent Length on Regio- and Enantioselectivity., *Adv. Synth. Catal.* 349 (2007) 2049-2053.
[13] Mirza I.A., Yachnin B.J., Wang S., Grosse S., Bergeron H., Imura A., Iwaki H., Hasegawa Y., Lau P.C., Berghuis A.M., Crystal structures of cyclohexanone monooxygenase reveal complex domain movements and a sliding cofactor, J Am Chem Soc 131 (2009) 8848-8854.
[14] Malito E., Alfieri A., Fraaije M.W., Mattevi A., Crystal structure of a Baeyer-Villiger monooxygenase, Proc Natl Acad Sci U S A 101 (2004) 13157-13162.
[15] Rossmann M.G., Adams M.J., Buehner M., Ford G.C., Hackert M.L., Liljas A., Rao S.T., Banaszak L.J., Hill E., Tsernoglou D., Webb L., Letter: Molecular symmetry axes and subunit interfaces in certain dehydrogenases, J Mol Biol 76 (1973) 533-537.
[16] Wierenga R.K., Terpstra P., Hol W.G., Prediction of the occurrence of the ADP-

binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint, J Mol Biol 187 (1986) 101-107.
[17] Reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Res 40 D71-75.
[18] Magrane M., Consortium U., UniProt Knowledgebase: a hub of integrated protein data, Database (Oxford) 2011 bar009.
[19] Fraaije M.W., Wu J., Heuts D.P., van Hellemond E.W., Spelberg J.H., Janssen D.B., Discovery of a thermostable Baeyer-Villiger monooxygenase by genome mining, Appl Microbiol Biotechnol 66 (2005) 393-400.
[20] Iwaki H., Wang S., Grosse S., Bergeron H., Nagahashi A., Lertvorachon J., Yang J., Konishi Y., Hasegawa Y., Lau P.C., Pseudomonad cyclopentadecanone monooxygenase displaying an uncommon spectrum of Baeyer-Villiger oxidations of cyclic ketones, Appl Environ Microbiol 72 (2006) 2707-2720.
[21] Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res 25 (1997) 3389-3402.
[22] Li W., Jaroszewski L., Godzik A., Clustering of highly homologous sequences to reduce the size of large protein databases, Bioinformatics 17 (2001) 282-283.
[23] Larkin M.A., Blackshields G., Brown N.P., Chenna R., McGettigan P.A., McWilliam H., Valentin F., Wallace I.M., Wilm A., Lopez R., Thompson J.D., Gibson T.J., Higgins D.G., Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947-2948.
[24] Edgar R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res 32 (2004) 1792-1797.
[25] Clamp M., Cuff J., Searle S.M., Barton G.J., The Jalview Java alignment editor, Bioinformatics 20 (2004) 426-427.
[26] Guindon S., Gascuel O., A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst Biol 52 (2003) 696-704.
[27] Huson D.H., Richter D.C., Rausch C., Dezulian T., Franz M., Rupp R., Dendroscope: An interactive viewer for large phylogenetic trees, BMC Bioinformatics 8 (2007) 460.
[28] Mayrose I., Graur D., Ben-Tal N., Pupko T., Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior, Mol Biol Evol 21 (2004) 1781-1791.
[29] Suzek B.E., Huang H., McGarvey P., Mazumder R., Wu C.H., UniRef: comprehensive and non-redundant UniProt reference clusters, Bioinformatics 23 (2007) 1282-1288.
[30] Eppink M.H., Schreuder H.A., Van Berkel W.J., Identification of a novel conserved sequence motif in flavoprotein hydroxylases with a putative dual function in FAD/NAD(P)H binding, Protein Sci 6 (1997) 2454-2458.
[31] Vallon O., New sequence motifs in flavoproteins: evidence for common ancestry and tools to predict structure, Proteins 38 (2000) 95-114.
[32] Fraaije M.W., Kamerbeek N.M., van Berkel W.J., Janssen D.B., Identification of a Baeyer-Villiger monooxygenase sequence motif, FEBS Lett 518 (2002) 43-47.
[33] Riebel A., Dudek H.M., de Gonzalo G., Stepniak P., Rychlewski L., Fraaije M.W., Expanding the set of rhodococcal Baeyer-Villiger monooxygenases by high-throughput cloning, expression and substrate screening, Appl Microbiol Biotechnol 95 1479-1489.
[34] Shortle D., Mutational studies of protein structures and their stabilities, Q Rev Biophys 25 (1992) 205-250.
[35] Lichtarge O., Bourne H.R., Cohen F.E., An evolutionary trace method defines binding surfaces common to protein families, J Mol Biol 257 (1996) 342-358.

[36] Gelly J.C., de Brevern A.G., Protein Peeling 3D: new tools for analyzing protein structures, Bioinformatics 27 (2011) 132-133.
[37] Gelly J.C., de Brevern A.G., Hazout S., 'Protein Peeling': an approach for splitting a 3D protein structure into compact fragments, Bioinformatics 22 (2006) 129-133.
[38] Gelly J.C., Etchebest C., Hazout S., de Brevern A.G., Protein Peeling 2: a web server to convert protein structures into series of protein units, Nucleic Acids Res 34 (2006) W75-78.
[39] Sowdhamini R., Blundell T.L., An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins, Protein Sci 4 (1995) 506-520.
[40] Liang J., Edelsbrunner H., Woodward C., Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design, Protein Sci 7 (1998) 1884-1897.
[41] Stierand K., Maass P.C., Rarey M., Molecular complexes at a glance: automated generation of two-dimensional complex diagrams, Bioinformatics 22 (2006) 1710-1716.
[42] Stierand K., Rarey M., From modeling to medicinal chemistry: automatic generation of two-dimensional complex diagrams, ChemMedChem 2 (2007) 853-860.

LEGENDS

Figures

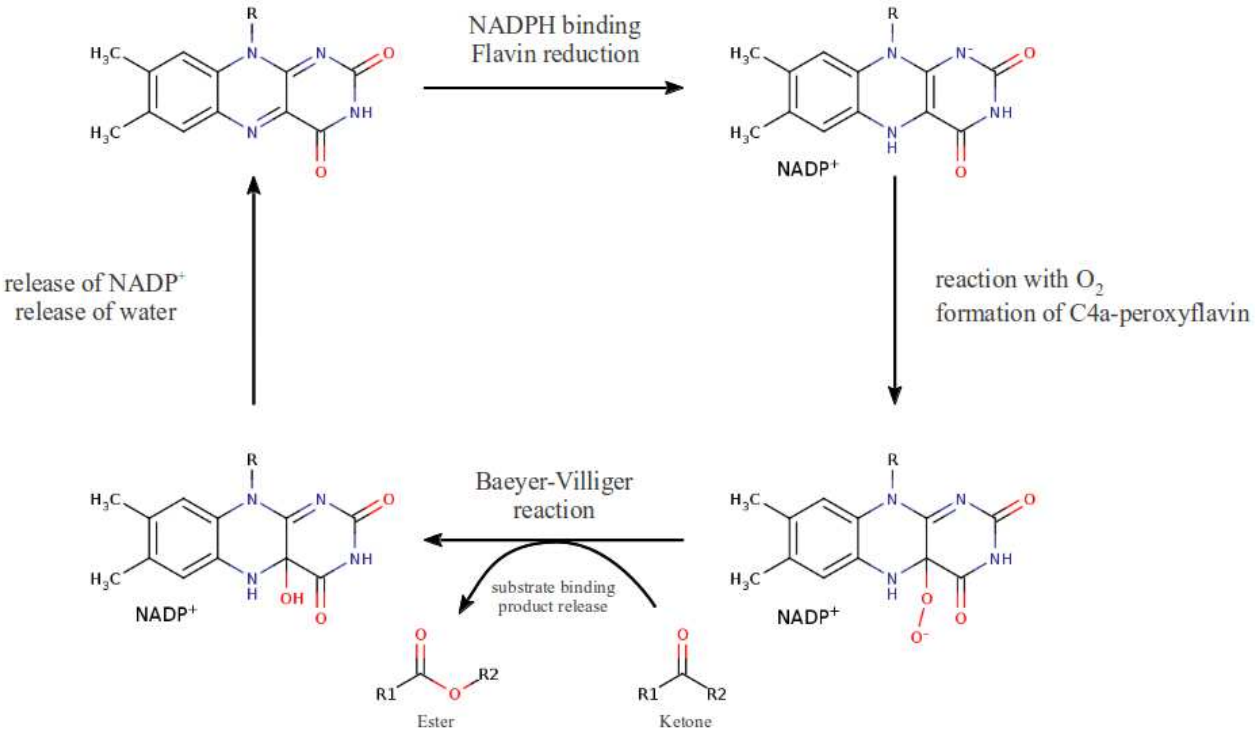

Figure 1: Simplified scheme of the catalytic mechanism of type I BVMOs.

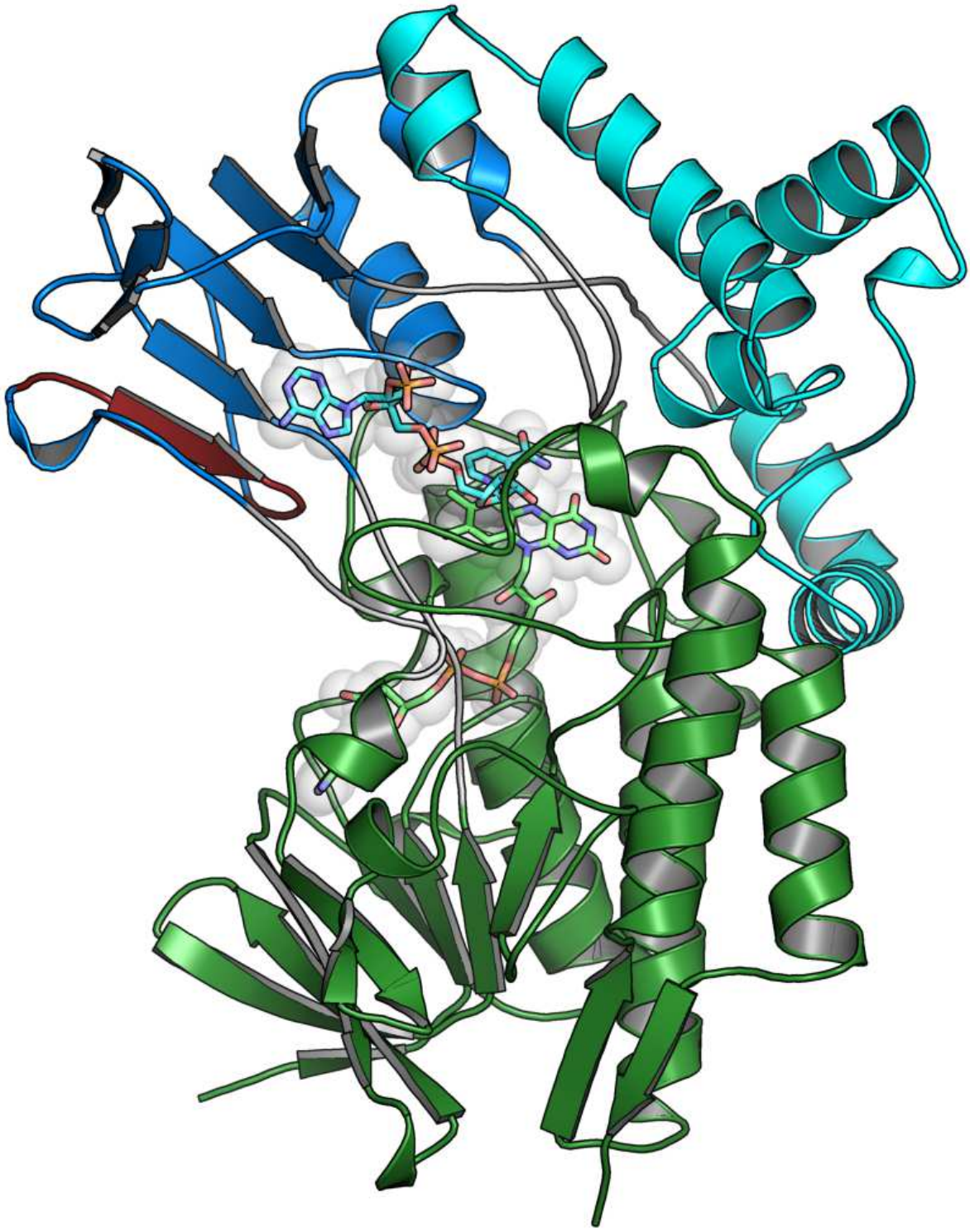

Figure 2: Structure of CHMO (PDB id 3GWD) [13]. FAD domain (residues 1 to 140 and 387 to 540) in green; NADPH domain (residue 152 to 208 and 335 to 380) in blue; helical domain (residues 224 to 322) in cyan. Linkers and mobile loops are shown in silver. BVMO signature motif is colored in red (residues 160 to 171). The visualization was done with PyMol software.

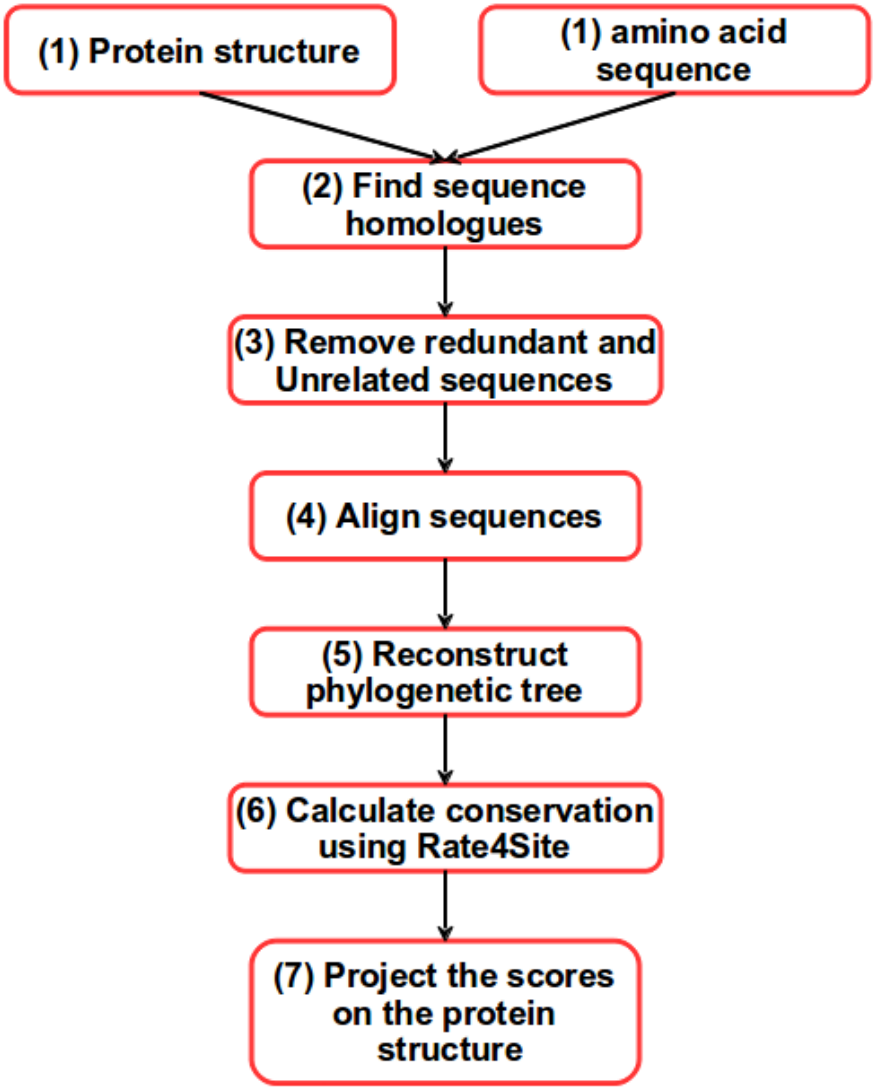

Figure 3: Flowchart resuming the global protocol used to look for BVMO similar sequences and to estimate the conservation score of residues during evolution. (1) from a limited set of annoted sequences, (2) a supervised dataset of related proteins was found using PSI-BLAST. (3) Only certain related sequences were selected. (4) They were aligned and (5) phylogeny was performed using maximum likelihood method. (6) conservation score for each position of the sequence was calculated and (7) analyzed on available protein structure.

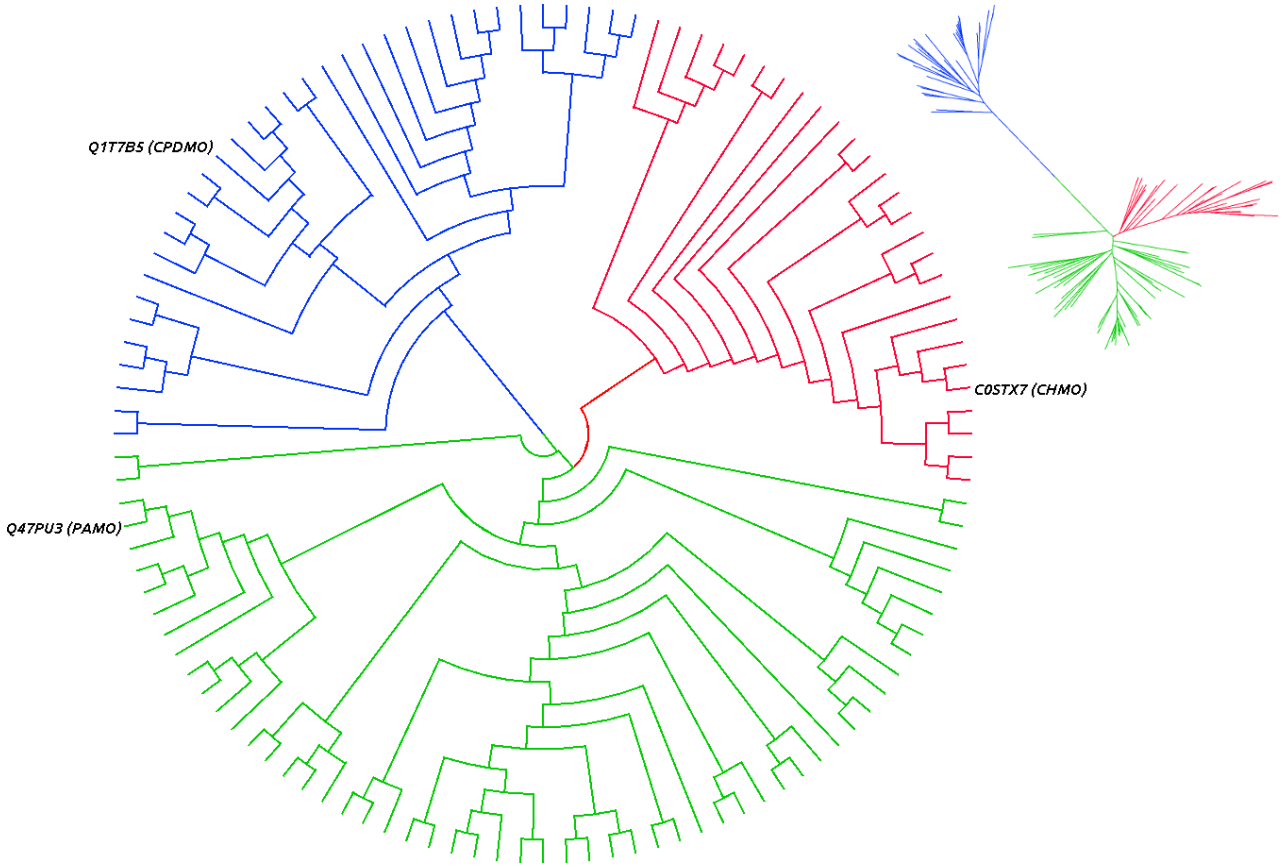

Figure 4: Circular cladogram presentation of the tree generated by PhyML [22] from the multiple sequence alignment obtained by MUSCLE [20]. Small radial phylogram is drawn on the top right to highlight evolutive distances between branches. The three initial sequences (CHMO, PAMO and CPDMO) are indicated and are enclosed in the red, green and blue clusters respectively.

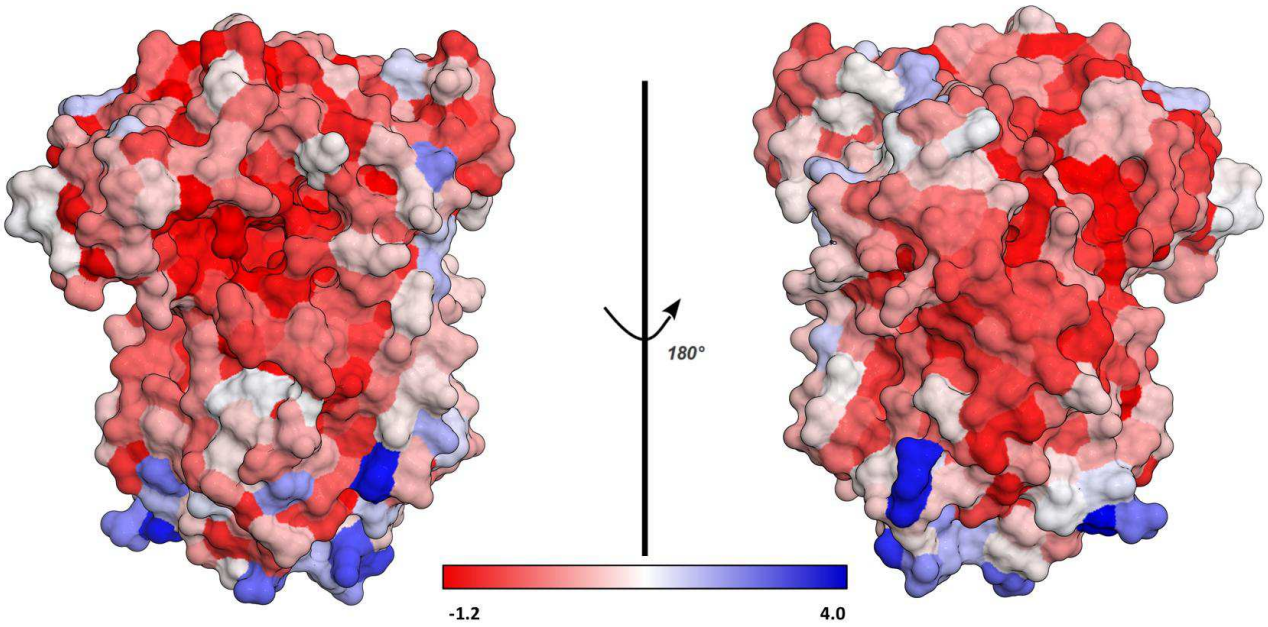

Figure 5: Amino acid conservation score seen on the protein structure of CHMO protein (PDB code 3GWD) [13] presented in surface (left) and another view after rotation along the Y axis (right). Red and blue colors correspond to high and low conserved residues respectively. Visualization done with PyMol software and the conservation score range is shown at the bottom of the figure.

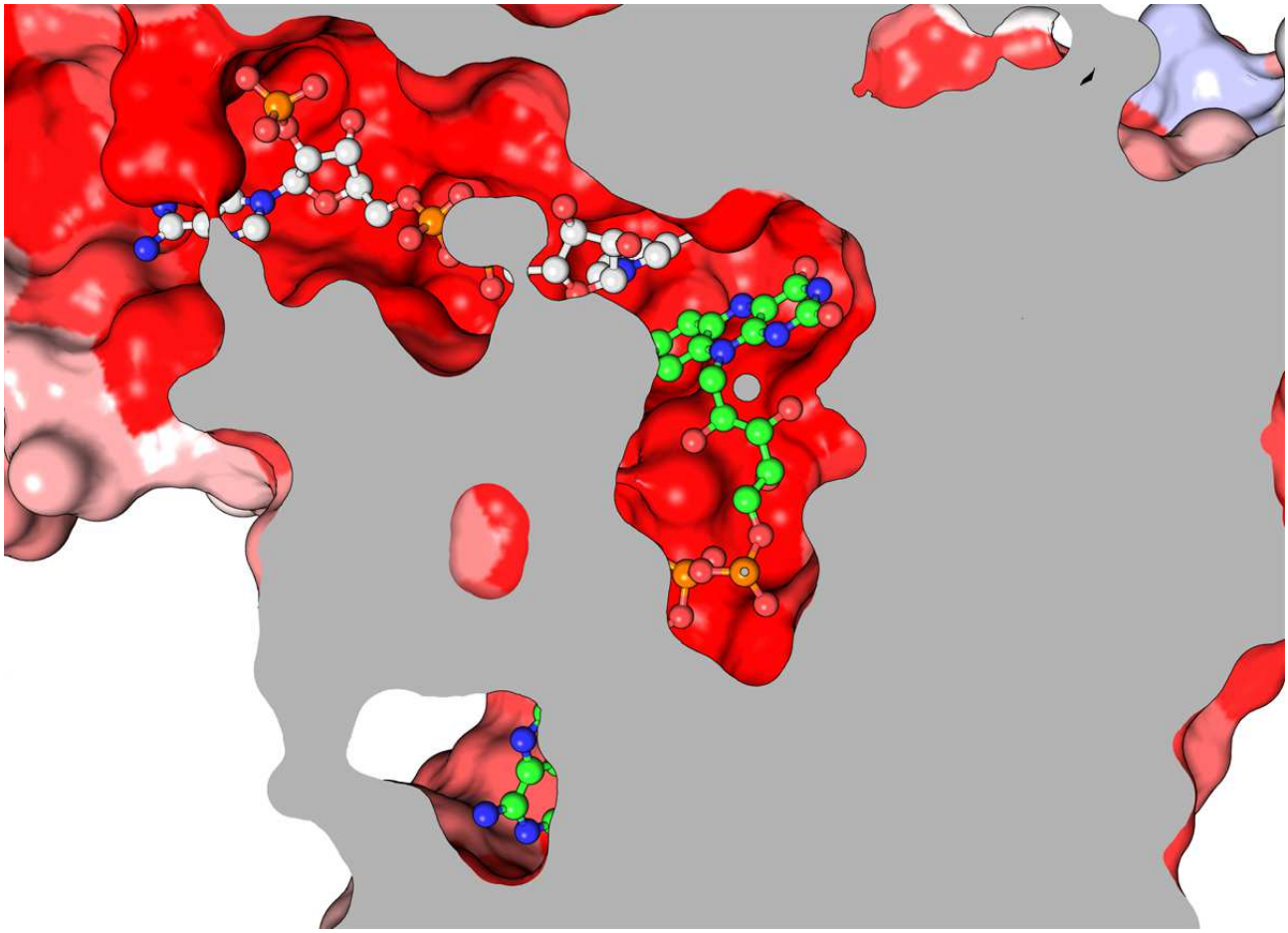

Figure 6: Analysis of FAD and NADPH cofactor domains. The protein surface of CHMO (PDB code 3GWD) is colored in term of sequence conservation while cofactors are shown in ball and stick representation and the carbon atoms of FAD and NADPH are colored in green and white respectively. Visualization done with PyMol software and the same conservation score color range was used as in Figure 5.

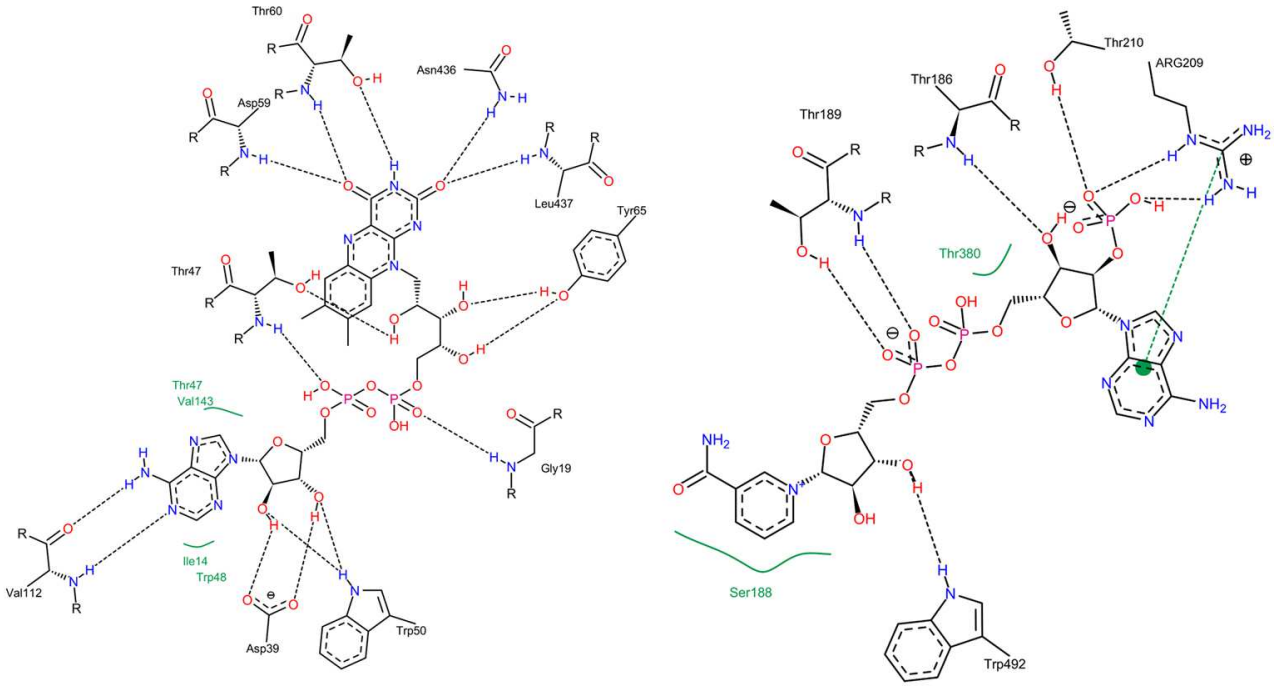

Figure 7: 2D representation of the interactions between CHMO (PDB code 3GWD) [13] and FAD (left) and NADPH(right) cofactors. Picture was generated by PoseViewWeb 1.97.0 [39, 40].

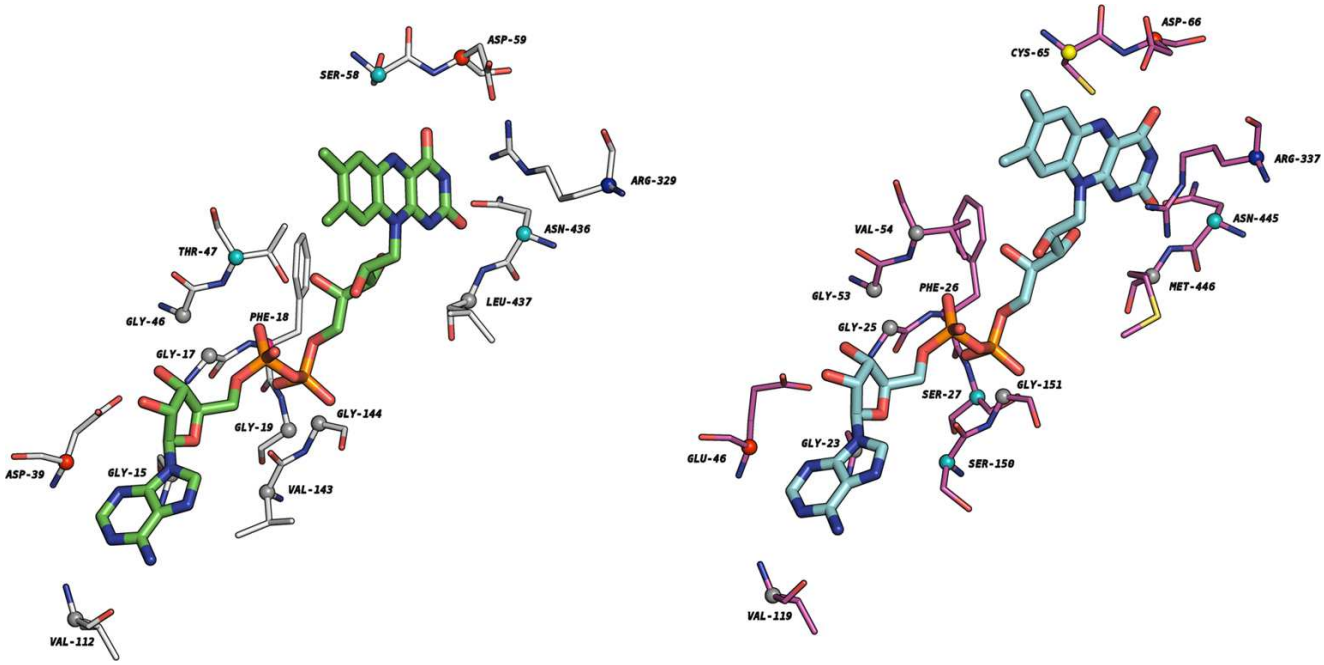

Figure 8: Binding site comparison between CHMO from *Rhodococcus* sp. HI-31 (PDB code: 3GWD - left) [13] and PAMO from *Thermofibidafusca* (PDB code 1W4X - right) [14]. FAD cofactor was shown in large stick presentation in green and cyan respectively. Side chains around cofactor are labeled. Colored spheres were positioned of the Cα atom depending on the residue type: Hydrophobic in gray; aromatic in pink; polar in light blue; positive in blue; negative in red and cysteine in yellow. Visualization done with PyMol software.

Table 1: Summary of the most conserved residues with their position in regards to CHMO protein from Rhodococcus *sp. HI-31, with their percentage of conservation for the different motifs of great interests: Rossmann fold motifs (yellow), Type I BVMO signature motif (red) and at the FAD (green) and NADPH (blue) binding sites.*

Rossmann Fold motif		**Signature fingerprint sequence**		**FAD-binding domain**		**NADPH-binding domain**	
Residue	***(%)***	***Residue***	***(%)***	***Residue***	***(%)***	***Residue***	***(%)***
G 15	99.1	F 160	94.8	G/A/S 19	47.0 / 42.6 / 9.6	P 152	96.6
G 17	99.1	G 162	100.0	D/E 39	20.0 / 80.0	I 184	94.0
G/A 20	97.4 / 2.6	H 166	94.8	G 45	99.1	T 186	94.8%
G 185	100.0	T 167	93.1	G 46	100.0	T/S 189	50.9 / 49.1
G 187	100.0	W 170	100.0	T 47	91.3	R 209	100.0
G/A 190	62.1 / 35.3	P/D 171	70.7 / 28.4	W 48	100.0	T/S 210	81.9 / 15.5
				W/F/Y 50	80.9 / 11.3 / 4.3	R 329	100.0
				N 51	100.0	A 379	96.6
				Y 53	97.4	G 381	100.0
				D 59	100.0	W/Y 492	69.8 / 28.4
				Y 65	100.0		
				V 112	82.8		
				N 436	98.3		

Table 2: Listing of the residues at the FAD domain of CHMO and their equivalent in PAMO.

CHMO	**PAMO**
R329	R337
N436	N445
D59	D66
L437	M446
S58	CYS65
F18	F26
G15	G23
G17	G25
G19	S27
G144	G151
T47	V54
G46	G53
V143	S150
D39	E46
V112	V119