3D modelling and functional annotation of the Pleurotus eryngii laccase 4 enzyme

Matteo Bolner

February 4, 2019

Abstract

In UniprotKB/TrEMBL the protein ERY4 is endowed with a poor automatical annotation as a laccase, inferred from other proteins of the laccase family. In this project, I would like to verify whether its automatic annotation can be supported by a more thoroughful search, according to the concept of protein family and to the comparison with a given template of the laccase family. In order to do this, I adopted a procedure in which functional annotation is done after the evaluation of the model of the protein on the basis of a given template of the family. The procedure of building by homology allowed me to demonstrate that ERY4 is indeed a laccase; my results indicate that the protein can be endowed with all the GO terms associated to the template protein.

1 Introduction

ERY4 is a gene that codifies for a protein of the fungus *Pleurotus eryngii* which, after automatic annotation of the transcript, is believed to be a laccase enzyme. Laccases are polyphenol oxidases, which belong to the family of blue multicopper oxidases. They can be involved in both the polymerization and the degradation of lignin and other similar molecules [1]. The active site of these proteins consists of four copper atoms: one located in the primary oxidation site (T1), and three others arranged in a trinuclear cluster, where oxygen is reduced (T2 and two T3s). From T1, the electrons are transferred to T2 and T3. Each copper in the trinuclear cluster is bound through coordination bonds to the imidazole side chains of histidin residues, while T1 is bound to two histidines and a cystein residue. The bond between residues and copper atoms coordinates their positioning in the active site; for a more detailed description see paper [1].

2 Methods

2.1 Data bases

The ERY4 target sequence is obtained from the UniProt database (release 2018_11)[2] ; the 1GYC template sequence and structure are obtained from the PDB file in RCSB PDB (18.12.18 release)[3]. The GO terms are downloaded from the UniProt entry associated to 1GYC. Information in relation to the protein domains and secondary structure is derived from InterPro (v.71, 8.11.18)[4].

2.2 Computational methods

The template protein is found after local alignment of the target (B0JDP9) with the UNIPROTKB_PDB database with BLASTP (2.7.1+)[5], using the parameters highlighted in Table 1.

Table 1: Parameters used to retrieve the template with BLAST

Program	BLASTP 2.7.1+
Target database	uniprotkb_pdb
E-Threshold	10
Matrix	Blosum62(Auto)
Filtering	None
Gapped	Yes
Hits	50

In order to compare the target and template sequences, I run LALIGN [6] with the parameters highlighted in Table 2.

Table 2: Parameters used to align the target and template sequences

Alignment method	Global
Number of reported sub-alignments	3
E-value treshold	10.0
Scoring matrix	BLOSUM50 (default)
Opening gap penalty	-12 (default)
Extending gap penalty	-2 (default)

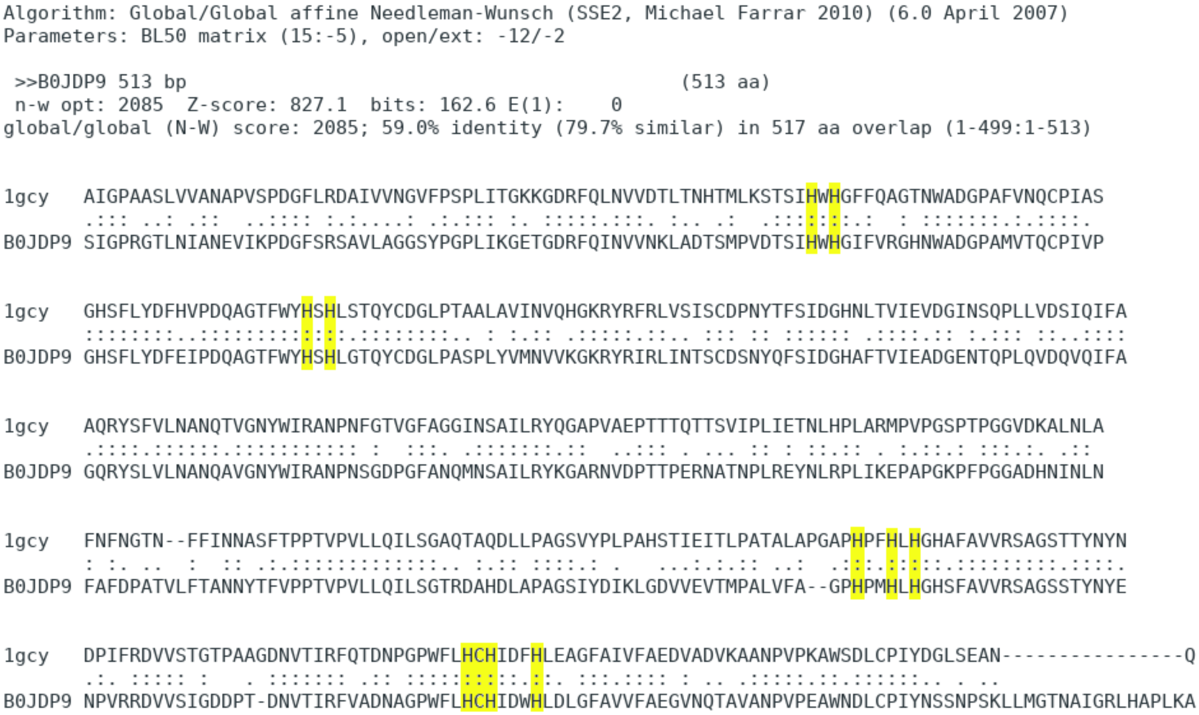
Modeller (release 9.20, r11208) is downloaded from https://salilab.org/modeller [7] and then used to generate three model structures of the target. Pymol is adapted as the molecular visualizer (Version 2.2.0)[8]. The online version of PROCHECK 3.5 [9] analyzes the stereochemical quality of the model. The computed target structure is aligned with the template structure with jCE (V2.11 2014 March 17)[10] installed locally on Manjaro linux 18.0.1 .

3 Template selection

The procedure of building by homology requires the identification of templates within the family. In order to locate the protein family, the first operation is to use BLASTP and locally align the target against UniprotKB_PDB. The results are filtered in order to show only the reviewed entries. A list of templates is obtained; the highest identity (61.9%) is with Q12718, corresponding to the PDB entry 1GYC [3] which was obtained with a resolution of 1.9 Å and manually annotated. The computed coverage is 97%, and the E-value is 0.0. 1GYC is therefore chosen as the best suiting template to represent the laccase enzyme in this project.

4 Sequence alignment

The signal peptide sequence was manually removed from the target sequence, since the protein in its active and mature form doesn't contain it. The template sequence was obtained from the PDB file of 1GYC. Figure 1 shows the alignment of the template with the target sequence. The sequence identity is 59% with a sequence alignment length of 513 residues. The histidine and cystein residues bound to copper atoms in the active site, described in the reference paper [1], are conserved, as highlighted in figure 1.



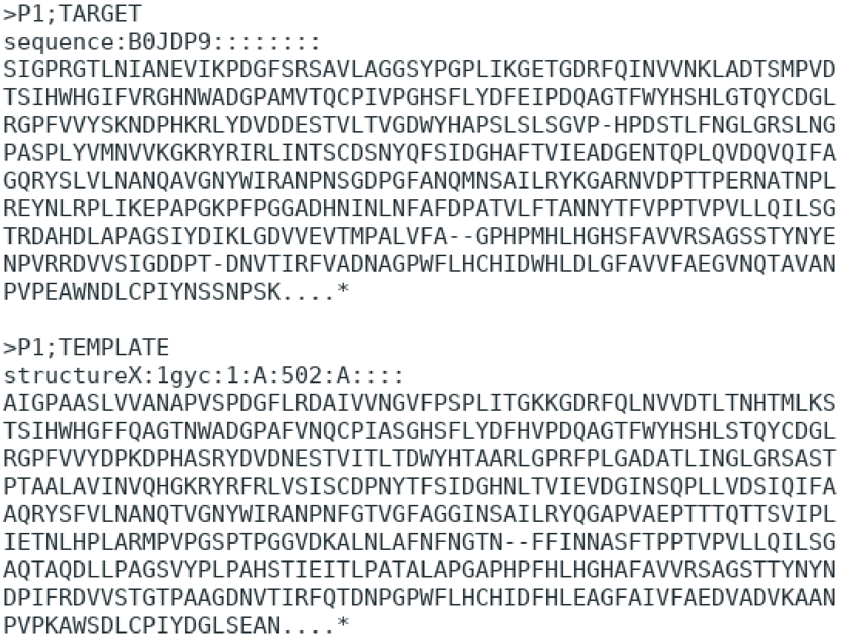
```
Algorithm: Global/Global affine Needleman-Wunsch (SSE2, Michael Farrar 2010) (6.0 April 2007)
Parameters: BL50 matrix (15:-5), open/ext: -12/-2

 >>B0JDP9 513 bp                                        (513 aa)
 n-w opt: 2085  Z-score: 827.1  bits: 162.6 E(1):    0
global/global (N-W) score: 2085; 59.0% identity (79.7% similar) in 517 aa overlap (1-499:1-513)


1gcy   AIGPAASLVVANAPVSPDGFLRDAIVVNGVFPSPLITGKKGDRFQLNVVDTLTNHTMLKSTSIHWHGFFQAGTNWADGPAFVNQCPIAS
       .::: ..: .::  ..:::: :.:....: .:.::: :. :::::.:::. :.. .:  .:::::::.:  : :::::::.:.::::.
B0JDP9 SIGPRGTLNIANEVIKPDGFSRSAVLAGGSYPGPLIKGETGDRFQINVVNKLADTSMPVDTSIHWHGIFVRGHNWADGPAMVTQCPIVP


1gcy   GHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLPTAALAVINVQHGKRYRFRLVSISCDPNYTFSIDGHNLTVIEVDGINSQPLLVDSIQIFA
       ::::::::..:::::::::::::.::::::::.. : :.:: .:::::.::.. ::: :: :::::: .::::.:: :.::: ::..::::
B0JDP9 GHSFLYDFEIPDQAGTFWYHSHLGTQYCDGLPASPLYVMNVVKGKRYRIRLINTSCDSNYQFSIDGHAFTVIEADGENTQPLQVDQVQIFA


1gcy   AQRYSFVLNANQTVGNYWIRANPNFGTVGFAGGINSAILRYQGAPVAEPTTTQTTSVIPLIETNLHPLARMPVPGSPTPGGVDKALNLA
       .::::.::::::.::::::::::: :  :::. .:::::::.::  ..::: . ... :: : ::.:: . :.::.: :::.:. .::
B0JDP9 GQRYSLVLNANQAVGNYWIRANPNSGDPGFANQMNSAILRYKGARNVDPTTPERNATNPLREYNLRPLIKEPAPGKPFPGGADHNINLN


1gcy   FNFNGTN--FFINNASFTPPTVPVLLQILSGAQTAQDLLPAGSVYPLPAHSTIEITLPATALAPGAPHPFHLHGHAFAVVRSAGSTTYNYN
       : :. ..  :  :: .:.:::::::::::::.. :.:: ::::.: .   ...:.:.:: ..:  .:::.:::::.:::::::::.::::.
B0JDP9 FAFDPATVLFTANNYTFVPPTVPVLLQILSGTRDAHDLAPAGSIYDIKLGDVVEVTMPALVFA--GPHPMHLHGHSFAVVRSAGSSTYNYE


1gcy   DPIFRDVVSTGTPAAGDNVTIRFQTDNPGPWFLHCHIDFHLEAGFAIVFAEDVADVKAANPVPKAWSDLCPIYDGLSEAN----------------Q
       .:. ::::: :   . ::::::: .:: ::::::::::.::. :::.:::: : .. .:::::.::.::::::.. . ..
B0JDP9 NPVRRDVVSIGDDPT-DNVTIRFVADNAGPWFLHCHIDWHLDLGFAVVFAEGVNQTAVANPVPEAWNDLCPIYNSSNPSKLLMGTNAIGRLHAPLKA
```

Figure 1

5 Modeller at work

5.1 Input preparation

The global alignment derived from Lalign was converted into .pir format as seen in figure 2:



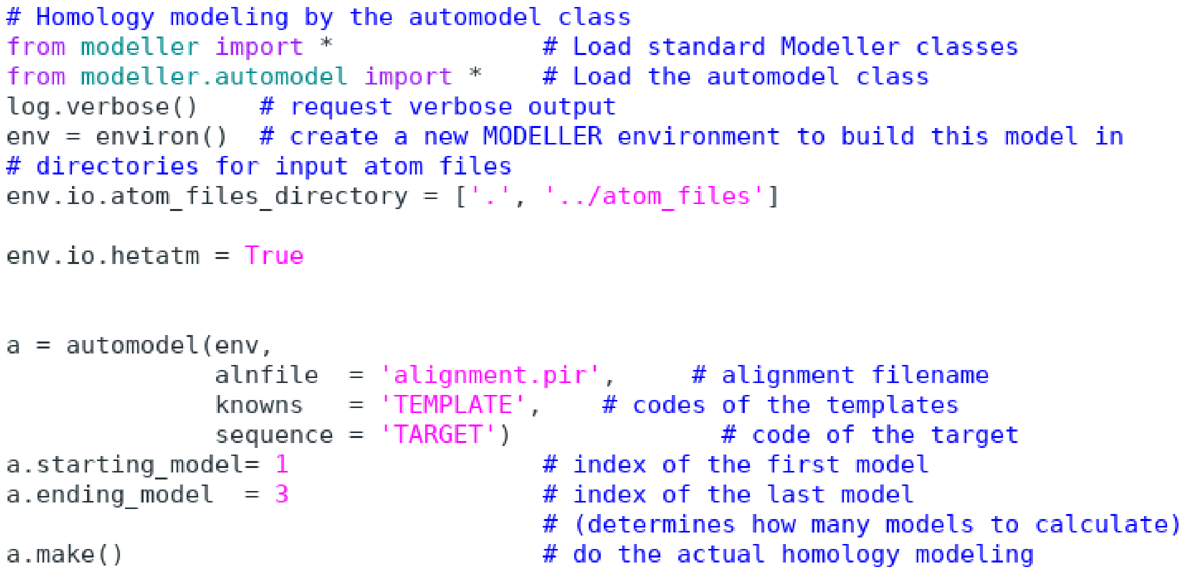
```
>P1;TARGET
sequence:B0JDP9::::::::
SIGPRGTLNIANEVIKPDGFSRSAVLAGGSYPGPLIKGETGDRFQINVVNKLADTSMPVD
TSIHWHGIFVRGHNWADGPAMVTQCPIVPGHSFLYDFEIPDQAGTFWYHSHLGTQYCDGL
RGPFVVYSKNDPHKRLYDVDDESTVLTVGDWYHAPSLSLSGVP-HPDSTLFNGLGRSLNG
PASPLYVMNVVKGKRYRIRLINTSCDSNYQFSIDGHAFTVIEADGENTQPLQVDQVQIFA
GQRYSLVLNANQAVGNYWIRANPNSGDPGFANQMNSAILRYKGARNVDPTTPERNATNPL
REYNLRPLIKEPAPGKPFPGGADHNINLNFAFDPATVLFTANNYTFVPPTVPVLLQILSG
TRDAHDLAPAGSIYDIKLGDVVEVTMPALVFA--GPHPMHLHGHSFAVVRSAGSSTYNYE
NPVRRDVVSIGDDPT-DNVTIRFVADNAGPWFLHCHIDWHLDLGFAVVFAEGVNQTAVAN
PVPEAWNDLCPIYNSSNPSK....*

>P1;TEMPLATE
structureX:1gyc:1:A:502:A::::
AIGPAASLVVANAPVSPDGFLRDAIVVNGVFPSPLITGKKGDRFQLNVVDTLTNHTMLKS
TSIHWHGFFQAGTNWADGPAFVNQCPIASGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGL
RGPFVVYDPKDPHASRYDVDNESTVITLTDWYHTAARLGPRFPLGADATLINGLGRSAST
PTAALAVINVQHGKRYRFRLVSISCDPNYTFSIDGHNLTVIEVDGINSQPLLVDSIQIFA
AQRYSFVLNANQTVGNYWIRANPNFGTVGFAGGINSAILRYQGAPVAEPTTTQTTSVIPL
IETNLHPLARMPVPGSPTPGGVDKALNLAFNFNGTN--FFINNASFTPPTVPVLLQILSG
AQTAQDLLPAGSVYPLPAHSTIEITLPATALAPGAPHPFHLHGHAFAVVRSAGSTTYNYN
DPIFRDVVSTGTPAAGDNVTIRFQTDNPGPWFLHCHIDFHLEAGFAIVFAEDVADVKAAN
PVPKAWSDLCPIYDGLSEAN....*
```

Figure 2: Lalign alignment converted into .pir format

The second line of both template and target sequences was modified as follows:

- field 1 contains a specification of whether or not 3D structure is available (structureX for a structure obtained via x-ray diffraction, and sequence if there is no structure available)
- field 2 contains the path to the file
- field 3 and 4 contain respectively the position of the first residue in the sequence and the chain it belongs to(to be specified only for the template)
- field 5 and 6 contain respectively the position of the last residue in the sequence and the chain it belongs to (as the 3rd and 4th field, specified only in the template) The gap at the end of the alignment was removed.

The sequences were both modified by adding four "." characters in the end, in order to allow the subsequent inclusion of four copper atoms; the end of the sequences was marked by the "*" character. Once the .pir file was obtained, the model.py script was modified in order to run Modeller (Figure 3) ; in particular, the names of the alignment file and the name of the template and target were specified. The number of models to produce was set to 3.



```
# Homology modeling by the automodel class
from modeller import *              # Load standard Modeller classes
from modeller.automodel import *    # Load the automodel class
log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in
# directories for input atom files
env.io.atom_files_directory = ['.', '../atom_files']

env.io.hetatm = True


a = automodel(env,
              alnfile  = 'alignment.pir',     # alignment filename
              knowns   = 'TEMPLATE',    # codes of the templates
              sequence = 'TARGET')              # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 3                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modeling
```

Figure 3: model.py script

5.2 Modeller output

Table 3 lists the molpdf (molecular probability density function) scores of the three different models required. The scores are computed according to Modeller's internal validation procedure. TARGET.B99990002, having the lowest molpdf score is adopted as the final result for modelling the structure of the target.

Filename	**molpdf**
TARGET.B99990001.pdb	3602.29688
TARGET.B99990002.pdb	3328.70459
TARGET.B99990003.pdb	3850.48364

Table 3: Modeller output table

The model stability was also evaluated with PROCHECK. According to PROCHECK, based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions. As shown in the ramachandran plot (Figure 4) the model is stable, with 91.5% of the residues having the proper torsion angles and being accordingly in the most favoured regions of the main chain backbone.

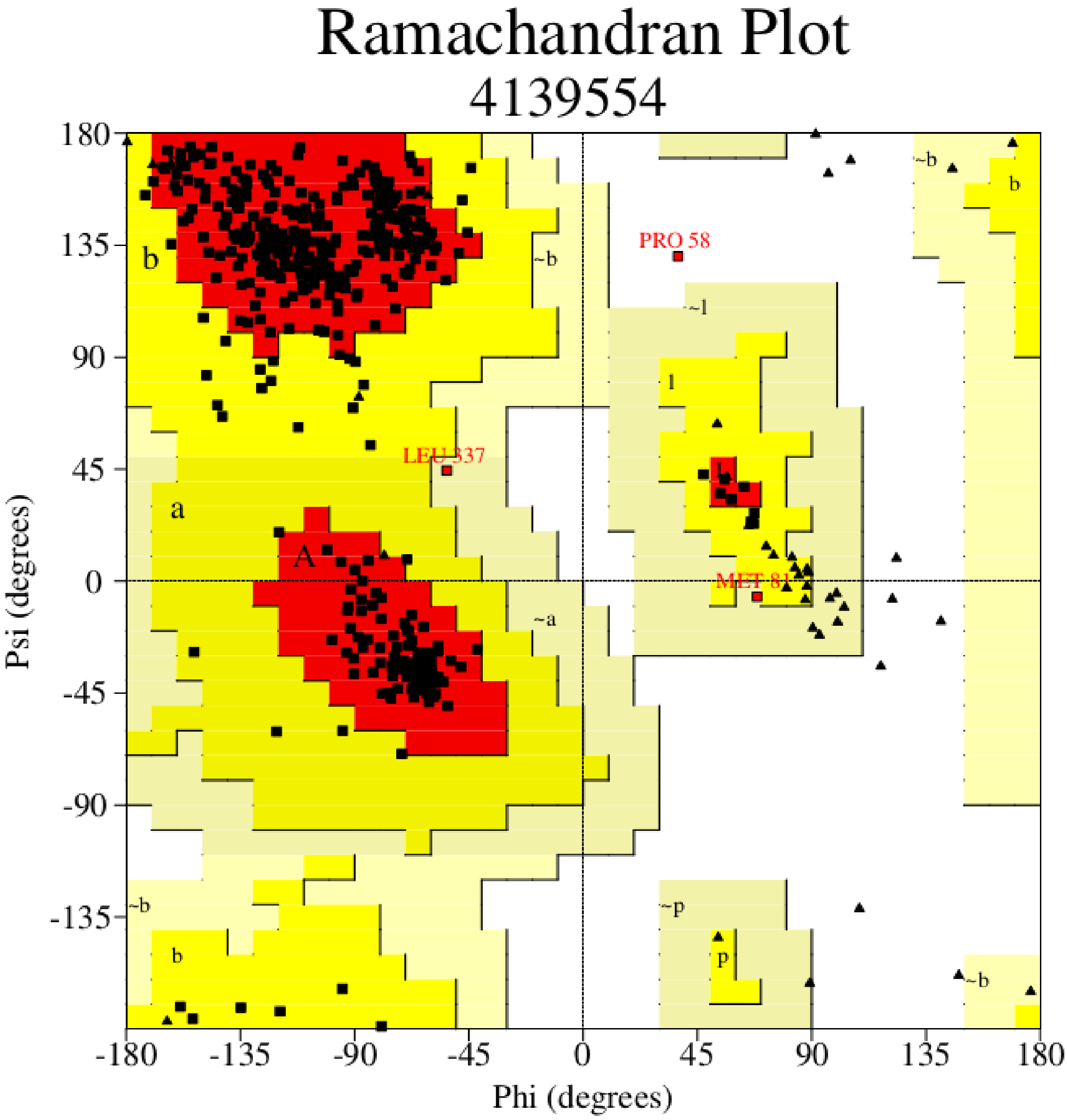


Figure 4: Ramachandran plot of model 2

Residues in most favoured regions [A,B,L]	375	91.5%
Residues in additional allowed regions [a,b,l,p]	33	8.0%
Residues in generously allowed regions [∼a,∼b,∼l,∼p]	2	0.5%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	410	100%
Number of end-residues (excl. Gly and Pro)	6	
Number of glycine residues (shown as triangles)	42	
Number of proline residues	42	
Total number of residues	500	

Table 4: Ramachandran plot statistics

6 Target annotation

The template and model pdb structures were aligned with jCE; the resulting pdb file was analyzed using Pymol.

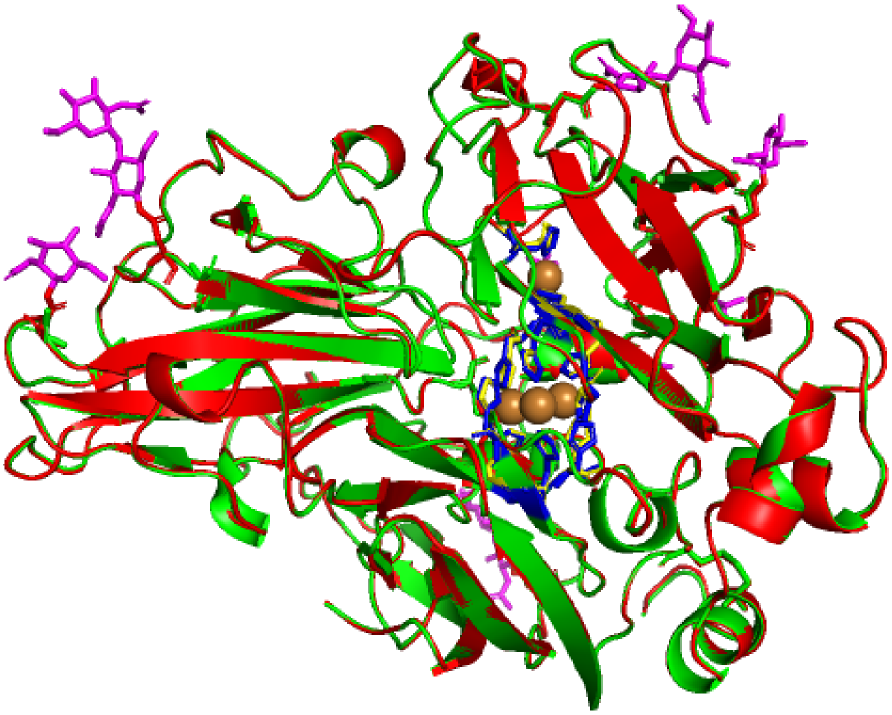
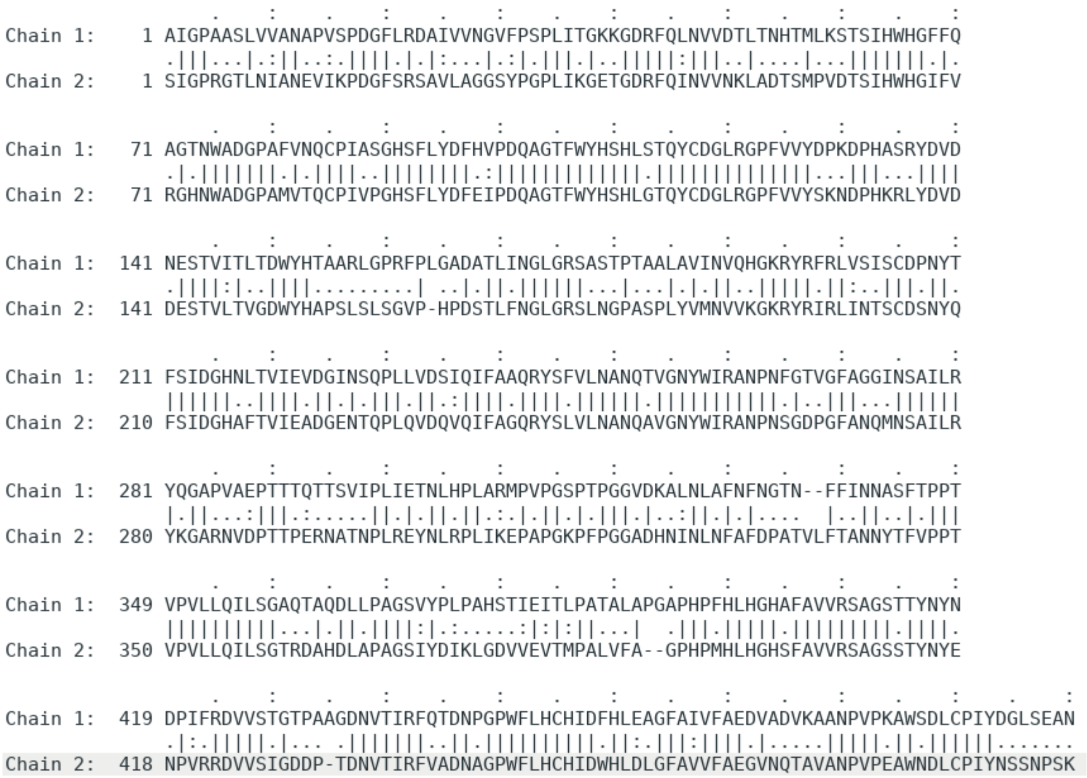


Figure 5: Structural superimposition of template(red) and model(green), the coppers can be seen as brown spheres in the active site, which is colored in blue; the N-acetylglucosamine are magenta coloured.

Figure 5 shows the superimposition of the template structure with the target model; the figure indicates that secondary structural motives are conserved. The RMSD is 0.26 A.



```
                  .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:    1 AIGPAASLVVANAPVSPDGFLRDAIVVNGVFPSPLITGKKGDRFQLNVVDTLTNHTMLKSTSIHWHGFFQ
              .|||...|.:||..:.||||.|.|:...|.:|.|||.|..|||||:|||..|....|...|||||||.|.
Chain 2:    1 SIGPRGTLNIANEVIKPDGFSRSAVLAGGSYPGPLIKGETGDRFQINVVNKLADTSMPVDTSIHWHGIFV

                  .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:   71 AGTNWADGPAFVNQCPIASGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPKDPHASRYDVD
              .|.|||||||.|.||||..||||||||.:|||||||||||||.||||||||||||||...|||...||||
Chain 2:   71 RGHNWADGPAMVTQCPIVPGHSFLYDFEIPDQAGTFWYHSHLGTQYCDGLRGPFVVYSKNDPHKRLYDVD

                  .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:  141 NESTVITLTDWYHTAARLGPRFPLGADATLINGLGRSASTPTAALAVINVQHGKRYRFRLVSISCDPNYT
              .||||:|..||||.........| ..|.||.||||||...|...|.|.||..|||||.||:..|||.||.
Chain 2:  141 DESTVLTVGDWYHAPSLSLSGVP-HPDSTLFNGLGRSLNGPASPLYVMNVVKGKRYRIRLINTSCDSNYQ

                  .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:  211 FSIDGHNLTVIEVDGINSQPLLVDSIQIFAAQRYSFVLNANQTVGNYWIRANPNFGTVGFAGGINSAILR
              ||||||..||||.||.|.|||.||.:||||.||||.||||||.|||||||||||.|..|||...||||||
Chain 2:  210 FSIDGHAFTVIEADGENTQPLQVDQVQIFAGQRYSLVLNANQAVGNYWIRANPNSGDPGFANQMNSAILR

                  .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:  281 YQGAPVAEPTTTQTTSVIPLIETNLHPLARMPVPGSPTPGGVDKALNLAFNFNGTN--FFINNASFTPPT
              |.||...:|||.:.....||.|.||.||.:.|.||.|.|||.|..:||.|.|....  |..||..|.|||
Chain 2:  280 YKGARNVDPTTPERNATNPLREYNLRPLIKEPAPGKPFPGGADHNINLNFAFDPATVLFTANNYTFVPPT

                  .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:  349 VPVLLQILSGAQTAQDLLPAGSVYPLPAHSTIEITLPATALAPGAPHPFHLHGHAFAVVRSAGSTTYNYN
              ||||||||||...|.||.||||:|.:.....:|:|:||...|  .|||.|||||.|||||||||.||||.
Chain 2:  350 VPVLLQILSGTRDAHDLAPAGSIYDIKLGDVVEVTMPALVFA--GPHPMHLHGHSFAVVRSAGSSTYNYE

                  .    :    .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:  419 DPIFRDVVSTGTPAAGDNVTIRFQTDNPGPWFLHCHIDFHLEAGFAIVFAEDVADVKAANPVPKAWSDLCPIYDGLSEAN
              .|:.|||||.|... .|||||||..||.||||||||||.||:.|||:||||.|.....|||||.||.||||||.......
Chain 2:  418 NPVRRDVVSIGDDP-TDNVTIRFVADNAGPWFLHCHIDWHLDLGFAVVFAEGVNQTAVANPVPEAWNDLCPIYNSSNPSK
```

Figure 6: Sequence alignment of template(red) and model(green) derived from structural superimposition by jCE

Figure 6 shows the sequence alignment derived with jCE from the structural superimposition. The structure of the template compares fairly well with the structure of the model: the sequence identity as derived after structural alignment is 62%.

In order to verify whether the active site is conserved, in figures 7 and 8 the structural superimposition is magnified to the point in which the superimposition of the active sites is visible.

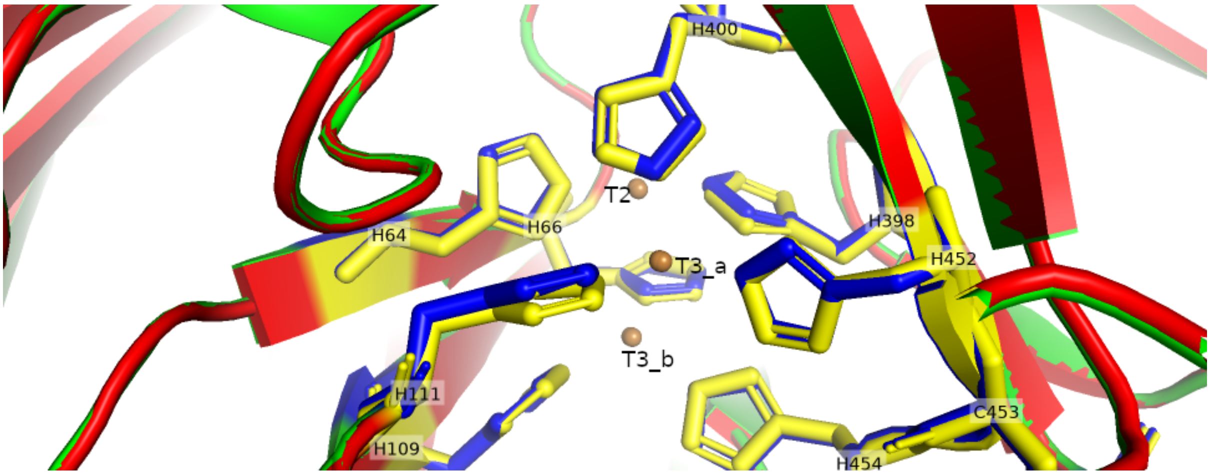


Figure 7: Structural superimposition of the active sites of template(red and yellow) and model(green and blue)

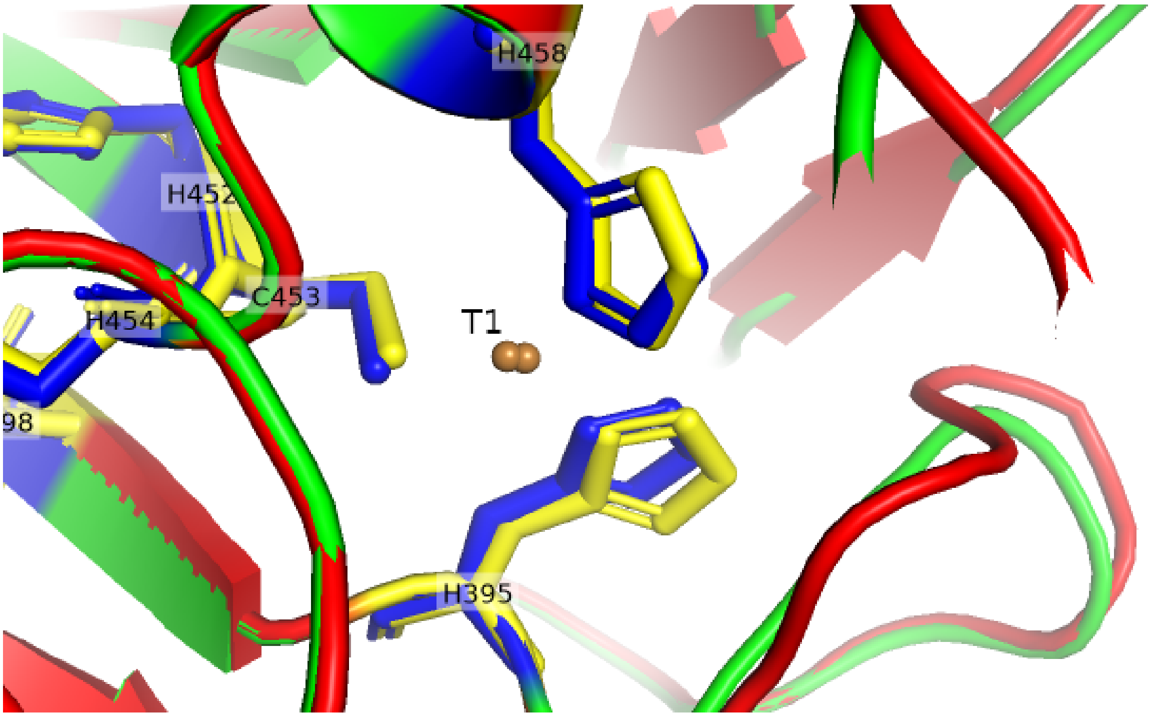


Figure 8: Structural superimposition of the active sites of template(red and yellow) and model(green and blue)

TEMPLATE			**MODEL**		
T2	T3_a	3.8	T2	T3_a	3.8
T2	T3_b	3.8	T2	T3_b	3.8
T3_a	T3_b	3.9	T3_a	T3_b	3.9
T1	H458	2.0	T1	H456	2.0
T1	H395	2.0	T1	H394	2.0
T1	C453	2.2	T1	C451	2.2
T3_a	H452	2.2	T3_a	H450	2.2
T3_a	H111	2.2	T3_a	H111	2.3
T3_a	H400	2.1	T3_a	H399	2.1
T3_b	H454	2.2	T3_b	H452	2.2
T3_b	H109	2.1	T3_b	H109	2.1
T3_b	H66	2.2	T3_b	H66	2.2
T2	H64	2.0	T2	H64	2.0
T2	H398	2.0	T2	H397	2.0

Table 5: Distances in the active sites of template and model (measured in Å)

Adopting the measurement tool in Pymol, the relative distances in the active sites of both target and model were obtained, and are shown in table 5. In agreement with what was observed in the reference paper [1], all distances measured are within 0.1 Å of difference when compared between template and model. From these results, I can obtain a confirmation that the active site is conserved in the computed model. The two disulphide bridges between Cys205 and Cys117 and between Cys85 and Cys488 are also conserved, as shown in figure 9: the bond's influence on the protein stability is therefore conserved.

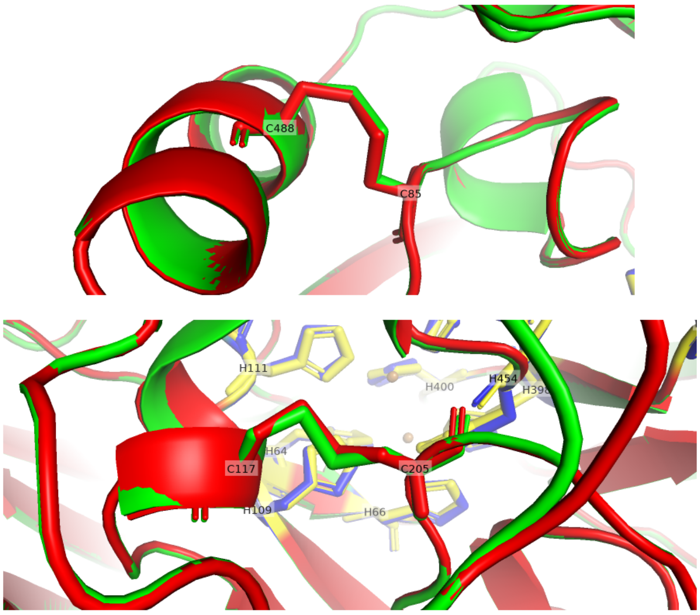


Figure 9: Superimposition of cysteins forming a disulphide bond in template(red) and model(green)

In addition, the conservation of the glycosilation sites described on PDB (https://www.rcsb.org/pdb/explore/macroMoleculeData.do?structureId=1GYC) and in the reference article was investigated: as shown in figure 10 , only three over six of the asparagins are conserved (Figure 10 A,E,F). As reported in the reference article, the conservation of only three glycosilation sites should not have a significant impact on the stability of the model.

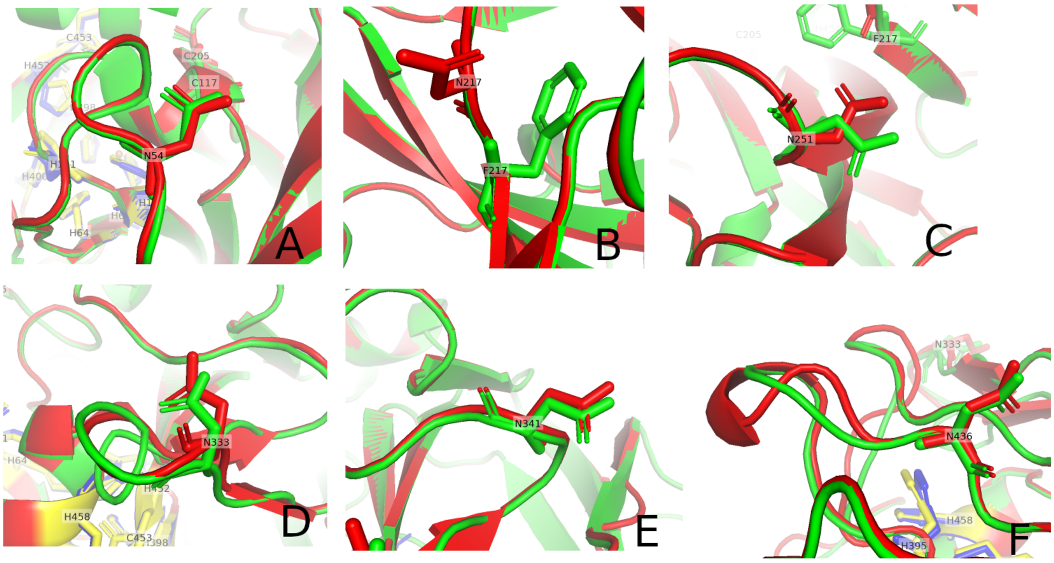
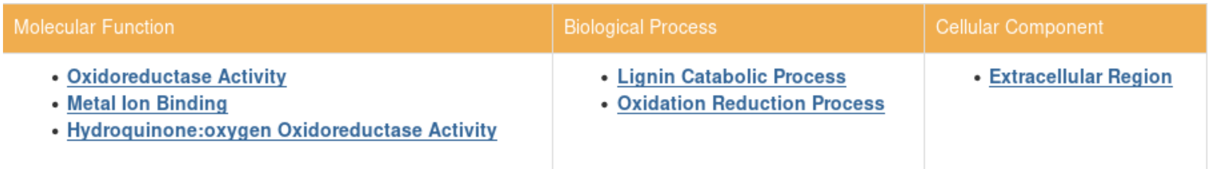


Figure 10: Superimposition of the glycosilation sites of template(red) and model(green)

Having confirmed the conservation of the active site, and having previously obtained the confirmation that the model is a stable protein, I can infer that the model protein is an active protein; therefore I have all the necessary information to predict that the protein can function as a laccase.

7 Discussion

Through the computation of a model structure, I have shown how the ERY4 sequence can be functionally annotated as a laccase and therefore, considering the GO terms associated to the template (figure 10), they may be transferred to our target.



Molecular Function	Biological Process	Cellular Component
• Oxidoreductase Activity • Metal Ion Binding • Hydroquinone:oxygen Oxidoreductase Activity	• Lignin Catabolic Process • Oxidation Reduction Process	• Extracellular Region

Figure 11: GO terms associated to the template (image taken from https://www.rcsb.org/pdb/explore/macroMoleculeData.do?structureId=1GYC)

8 References

1 Piontek, K., Antorini, M., & Choinowski, T. (2002, October 04). "Crystal structure of a laccase from the fungus Trametes versicolor at 1.90-A resolution containing a full complement of coppers."

2 https://www.uniprot.org/uniprot/B0JDP9

- The UniProt Consortium "UniProt: the universal protein knowledgebase" Nucleic Acids Res. 46: 2699 (2018)

3 https://www.rcsb.org/structure/1gyc

- "The Protein Data Bank" H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) Nucleic Acids Research, 28: 235-242. doi:10.1093/nar/28.1.235

4 https://www.ebi.ac.uk/interpro/

- Alex L Mitchell, Teresa K Attwood, Patricia C Babbitt, Matthias Blum, Peer Bork, Alan Bridge, Shoshana D Brown, Hsin-Yu Chang, Sara El-Gebali, Matthew I Fraser, Julian Gough, David R Haft, Hongzhan Huang, Ivica Letunic, Rodrigo Lopez, Aurélien Luciani, Fabio Madeira, Aron Marchler-Bauer, Huaiyu Mi, Darren A Natale, Marco Necci, Gift Nuka, Christine Orengo, Arun P Pandurangan, Typhaine Paysan-Lafosse, Sebastien Pesseat, Simon C Potter, Matloob A Qureshi, Neil D Rawlings, Nicole Redaschi, Lorna J Richardson, Catherine Rivoire, Gustavo A Salazar, Amaia Sangrador-Vegas, Christian J A Sigrist, Ian Sillitoe, Granger G Sutton, Narmada Thanki, Paul D Thomas, Silvio C E Tosatto, Siew-Yit Yong and Robert D Finn (2019). "InterPro in 2019: improving coverage, classification and access to protein sequence annotations." Nucleic Acids Research, Jan 2019; doi: 10.1093/nar/gky1100

5 https://www.uniprot.org/blast

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed

6 https://embnet.vital-it.ch/software/LALIGN_form.html

- Huang, Xiaoqiu & Miller, Webb. (1991). "A Time-Efficient, Linear-Space Local Similarity Algorithm." Advances in Applied Mathematics. 12. 337-357.

7 https://salilab.org/modeller/

- A. Sali & T.L. Blundell. "Comparative protein modelling by satisfaction of spatial restraints." J. Mol. Biol. 234, 779-815, 1993

8 https://pymol.org

- DeLano, W. L. (2002). "Pymol: An open-source molecular graphics tool." CCP4 Newsletter On Protein Crystallography, 40, 82-92.

9 http://servicesn.mbi.ucla.edu/PROCHECK

- Laskowski R A, MacArthur M W, Moss D S, Thornton J M (1993). "PROCHECK - a program to check the stereochemical quality of protein structures." J. App. Cryst., 26, 283-291.

- Laskowski R A, Rullmannn J A, MacArthur M W, Kaptein R, Thornton J M (1996). "AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR". J Biomol NMR, 8, 477-486. [PubMed id: 9008363]

10 source.rcsb.org/jfatcatserver/

- Shindyalov, I.N. & Zhuang, Pelion. (1998). "Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path." Protein engineering. 11. 739-47