Functional annotation of the Laccase from *Pleurotus eryngii* with comparative modelling

Author

Becchi Tommaso
University of Bologna
Bioinformatic Master Degree Course

Introduction

The aim of this work is to obtain a functional annotation of the Laccase from *Pleurotus eryngii* (UniProt entry: B0JDP9).

In UniProt[1] this protein is unreviewed and poorly annotated (annotation score equal to 1 out of 5) with experimental evidence only at transcript level (existence of cDNA(s), RT-PCR or Northern blots data indicate the existence of a transcript[2]). Therefore, it is necessary to individuate a good annotated protein which is similar to the target. After modelling and statistical validation procedures, we infer that it's possible to transfer GO terms for the annotation of the target protein.

Laccases (benzenediol oxygen oxidoreductase, ENZYME entry: EC 1.10.3.2[3]) are polyphenol oxidases, which belongs to the family of blue multicopper oxidases (bmCuO)[4]. This family catalyses the one-electron oxidation of four reducing-substrate molecules concomitant with the four-electron reduction of molecular oxygen to water (Figure 1).
In fungi and plants, laccases have a biological role in catalysing the remodelling and the degradation of lignin.

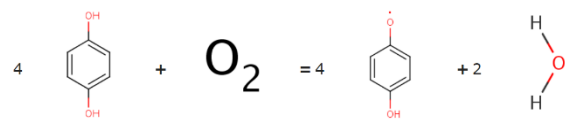

Figure 1: Catalytic activity of Laccases (ENZYME entry: EC 1.10.3.2). This image is obtained from Uniprot[1]

Blue multicopper oxidases contain one Type-1 (T1) copper, which is the primary oxidation site and at least three additional coppers: one Type-2 (T2) and two Type-3 (T3) coppers which are organised in a trinuclear cluster[4]. Consequently, the active site of the Laccases is characterized by the presence of these four copper atoms which are indispensable for catalytic activity. They are coordinated by conserved histidine and cysteine arranged so that the coppers are coplanar with the nitrogen and sulphur atoms that coordinate them.
This family has other important biological features: two disulphide bonds between two pairs of cysteine are present inside the molecule, furthermore there is a glycosylation pattern which involves several asparagines.
To annotate the target protein, I have to individuate all these conserved residues which have a specific biological role because if they are not present then it is not possible to transfer the information from the template.

Methods

Databases

The sequence and the annotation for both the target and the template are found in UniProt (release 2019_12)[1].
The PDB file with the atom coordinates of the template is obtained from RCSB PDB[5] (release 2019_12) .

Computational methods

BLASTp[6][7] (v. 2.9.0 as implemented in UniprotKB release 2019_12) was used for the template selection.
ExPASy LALIGN[8][9] was used to compute the global sequence alignment of the target against the template using Needleman and Wunsch algorithm.
Modeller[10][11] (v. 9.23) was used to built the models using building-by-homology procedures and implements comparative protein structure modeling by satisfaction of spatial restraints[12].
jCE[13] (v. 1.1) was used to check the quality of the models performing a structural superimposition. It obtains the protein structure alignment by incremental combinatorial extension[14] (ce) of the optimal path.
PROCHECK[15][16] (v.3.5) was used to generate the Ramachandran Plots of the models and then to evaluate their quality.

At the end RasMol[17] (v. 2.7.5.2) is used to observe the conformation of the active site both in the template and in the model to verify the conservations of the main residues and their distances. RasMol is a molecular visualization software with allows to manipulate the observed molecules to investigate them.

Template selection

A BLAST[18] search against "3D structure (PDB)" database was performed by adopting the following parameters:

Matrix	Blosum62
Threshold	10
Filtered	False
Gapped	True
Maximum number of hits reported	250

I obtain 70 results: 53 are unreviewed entry and only 17 are reviewed. The following criteria were adopted to choose the best template: it must be a reviewed protein with a good annotation, it is necessary a high sequence identity with the target (at least 30%) and high coverage of the target. It is also required the availability of a high-resolution crystal structure of the protein.

Laccase-2 of *Trametes versicolor* (UniProt entry: Q12718) is the best result because it satisfies all the conditions explained above: it is the second best result for similarity, it is reviewed (annotation score is equal to 4 out of 5) with experimental evidence at protein level. Furthermore, in the PDB database an X-ray structure (with a resolution of 1.9 Å) of this molecule is present (PDB entry: 1GYC).

Observing the sequence which derives from the PDB structure it is possible to notice that there is a main difference with the sequence in UniProt: in this sequence the 20 initial amino acids are missing because in the X-ray structure there is no signal peptide.

Analyzing the annotations in UniProt for Laccase-2 of *Trametes versicolor* it can be observed that there are several residues with a biological role: 10 histidine and 1 cysteine are involved in the coordination of the four copper atoms responsible for the enzymatic activity (Figure 2). Moreover 4 cysteine participate to 2 disulfide bonds and 8 asparagines are glycosylated. These elements are fundamental for the correct folding and activity of the protein so in Table 1 all these features are summed up.

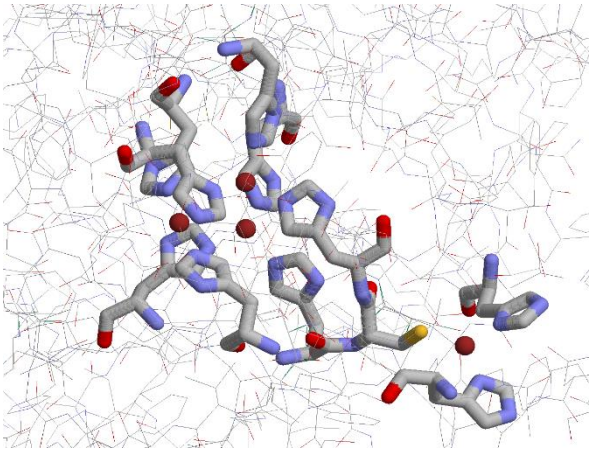

Figure 2: a 3D representation of the template active site is shown: the four copper atoms are shown in brown and the residues involved in their coordination are highlighted. The image is obtained using RasMol.

Position in Uniprot	Position in PDB	Residues	Info	Distance (Å)
84	64	his	Cu1	2,01
86	66	his	Cu2	2,15
129	109	his	Cu2	2,12
131	111	his	Cu3	2,23
415	395	his	Cu4	2,02
418	398	his	Cu1	1,97
420	400	his	Cu3	2,12
472	452	his	Cu3	2,16
473	453	cys	Cu4	2,19
474	454	his	Cu2	2,17
478	458	his	Cu4	2,04
105	85	cys	DB	1,98
508	488	cys		
137	117	cys	DB	2,04
225	205	cys		
74	54	asn	Glicosilation	1.42
161	141	asn	Glicosilation	*not available*
228	208	asn	Glicosilation	*not available*
237	217	asn	Glicosilation	1.41
271	251	asn	Glicosilation	1.40
353	333	asn	Glicosilation	1.39
361	341	asn	Glicosilation	1.43
456	436	asn	Glicosilation	1.42

Table 1: all the information about the most important residues in the Laccase-2 of *Trametes versicolor* are summed up in this table. The distances are obtained using RasMol. The difference between the position in UniProt sequence and the one in PDB sequence is due to the presence in the Uniprot entry of a signal peptide of 20 amino acids which is not present in the PDB structure.

Sequence alignment

LALIGN is used to compute the global alignment[19] between the target sequence obtained from UniProt (B0JDP9) and the template sequence obtained from PDB (1GYC). The parameters setting are the following: selected scoring matrix is BLOSUM50[20], gap opening and gap extension penalties of -12 and -2 respectively and E-value threshold of 10.0. The result of this alignment is visible in Figure S1 (Supplementary methods).

The sequences have an high identity percentage (56.8%) and they are very similar (76.7%).

Therefore, it possible to assert that this global alignment is useful for the subsequent creation of the model because the sequence identity is greater than 30% and consequently sequence similarity implies structural similarity and they probably belong to the same protein family[21].

Modeller

Modeller was run by using an alignment in PIR[22] format derived from the previously described alignment.

The obtained PIR file must be edited to allow a correct model generation then the following heteroatoms were also modelled: four coppers and eight NAGs.

The residues 1-20 and 496-513 were manually deleted from the target sequence because they show no pairing with the sequence of the template. Consequently, corresponding gap plus the residue Gln499 are removed from the template (all the atoms of Gln499 are also removed from the PDB file).

5 different models are generated using Modeller. Each of them is described by three parameters which are summed up in Table 2. The best model is the one which minimizes both the *molpdf* (molecular probability density function) value and the *DOPE score* (a parameter related to the energy of the molecule and its stability). According to these conditions the best model is Model 3.

	Molpdf	DOPE score	GA341 score
Model 1	3743.09	-57575.86	1.00
Model 2	3799.98	-57325.29	1.00
Model 3	3657.02	-57896.11	1.00
Model 4	3771.90	-57499.31	1.00
Model 5	3983.98	-57557.38	1.00

Table 2: This table shows the value of the 3 parameters for each obtained mode. These parameters allow to evaluate the quality of the model. *molpdf* is an objective function related to satisfaction of spatial restraints, *DOPE score* is related to the stability of the molecule and *GA341 score* is related to sequence identity

Subsequently it is necessary to investigate the spatial conformations of the atoms in the model then PROCHECK is used. PROCHECK uses the PDB file obtained with Modeller to generate a Ramachandran Plot for each model. Ramachandran plots allow to observe how many residues have torsion angles at the level of the backbone (*Psi* and *Phi*) in favoured regions or disallowed regions. Table 3 contains all these information and also with this analysis the best model is Model number 3.

	Favoured regions	Additional allowed regions	Generously allowed regions	Disallowed regions
Model 1	90,2%	9,5%	0,2%	0,0%
Model 2	89,8%	9,5%	0,7%	0,0%
Model 3	90,5%	9,3%	0,2%	0,0%
Model 4	89,3%	9,5%	1,0%	0,2%
Model 5	90,2%	9,3%	0,5%	0,0%

Table 3: For each model PROCHECK results are summarized. They are displayed as percentages of residues in the favoured, additionally allowed, generously allowed and disallowed regions. All the 5 models have similar results but the one with the best structural organisation is Model 3 which has the highest percentage of amino acids in "Favoured regions".

Therefore, after the observation of Modeller results and the validation with PROCHECK it is possible to assert that Model 3 is the best model and it can be used afterwards for the target annotation (Figure 4 shows its Ramachandran Plot).

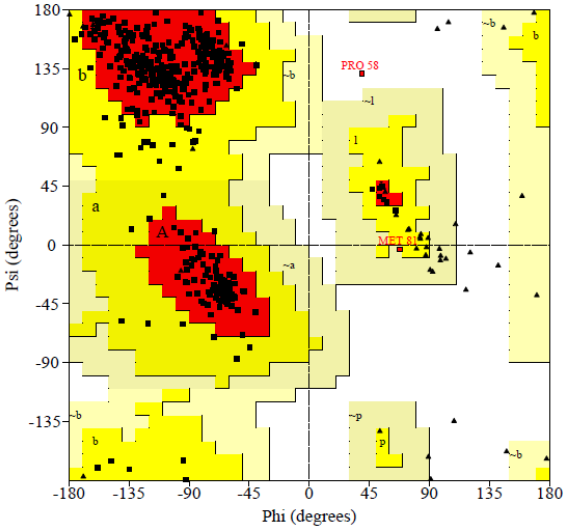

Figure 4: Ramachandran Plot obtained with PROCHECK of Model 3. In the red regions are present the residues with the most frequent torsion angles which identify α and β structures.

Target annotation

Model 3 is structurally superimposed to the template using jCE (Figure 5). The RMSD obtained with this structural alignment is 0.26 Å so it is possible to assert that the two structures superimpose very well not only at the level of the backbone but also in the lateral chains.

From the structural alignment it is possible to derivate sequence alignment (Figure 6). Comparison between sequence alignment derived from sequences alignment (Supplementary, Figure S1) and from structural superimposition (Figure 6) shows almost no differences because there is a very high percentage of identity between the two analysed sequnces (the main differences are due to the presence in the original sequences of the signal peptide).

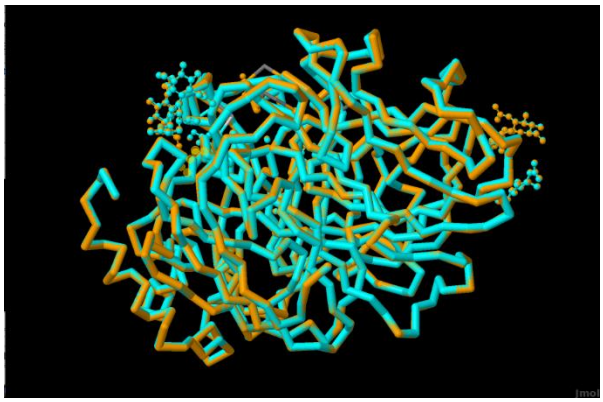

Figure 5: Structural superimposition with jCE between Model 3 and the template. The two structures have low RMSD and they superimpose very well both for the backbones and the lateral chains. The two sequences have different colours: one is orange and the other is blue, but we can see that they are almost always superimposed.

```
1:A       10:A      20:A      30:A      40:A      50:A      60:A      70:A
|         |         |         |         |         |         |         |
AIGPAASLVVANAPVSPDGFLRDAIVVNGVFPSPLITGKKGDRFQLNVVDTLTNHTMLKSTSIHWHGFFQ
.|||...|.:||.:.:||||.|.|:...|.:|.|||.|..|||||:|||..|....|...|||||||.|.
SIGPRGTLNIANEVIKPDGFSRSAVLAGGSYPGPLIKGETGDRFQINVVNKLADTSMPVDTSIHWHGIFV
|         |         |         |         |         |         |         |
1:        10:       20:       30:       40:       50:       60:       70:

          80:A      90:A      100:A     110:A     120:A     130:A     140:A
          |         |         |         |         |         |         |
AGTNWADGPAFVNQCPIASGHSFLYDFHVPDQAGTFWYHSHLSTQYCDGLRGPFVVYDPKDPHASRYDVD
.|.|||||||.|.||||..||||||||.:|||||||||||||.|||||||||||||...|||...||||
RGHNWADGPAMVTQCPIVPGHSFLYDFEIPDQAGTFWYHSHLGTQYCDGLRGPFVVYSKNDPHKRLYDVD
          |         |         |         |         |         |         |
          80:       90:       100:      110:      120:      130:      140:

          150:A     160:A     170:A     180:A     190:A     200:A     210:A
          |         |         |         |         |         |         |
NESTVITLTDWYHTAARLGPRFPLGADATLINGLGRSASTPTAALAVINVQHGKRYRFRLVSISCDPNYT
.||||:|..||||.........| ..|.|||.||||||...|...|.|.||..||||||.||:..|||.||.
DESTVLTVGDWYHAPSLSLSGVP-HPDSTLFNGLGRSLNGPASPLYVMNVVKGKRYRIRLINTSCDSNYQ
          |         |         |         |         |         |         |
          150:      160:      170:      180:      190:      200:

          220:A     230:A     240:A     250:A     260:A     270:A     280:A
          |         |         |         |         |         |         |
FSIDGHNLTVIEVDGINSQPLLVDSIQIFAAQRYSFVLNANQTVGNYWIRANPNFGTVGFAGGINSAILR
||||||..||||.||.|.|||.||.:|||||.||||.||||||.||||||||||.|..|||...||||||
FSIDGHAFTVIEADGENTQPLQVDQVQIFAGQRYSLVLNANQAVGNYWIRANPNSGDPGFANQMNSAILR
|         |         |         |         |         |         |         |
 210:      220:      230:      240:      250:      260:      270:

          290:A     300:A     310:A     320:A     330:A       340:A
          |         |         |         |         |         |         |
YQGAPVAEPTTTQTTSVIPLIETNLHPLARMPVPGSPTPGGVDKALNLAFNFNGTN--FFINNASFTPPT
|.||...:|||.:.....||.|.||.||.:.|.||.|.|||.|..:||.|.|.....  |..||..|.|||
YKGARNVDPTTPERNATNPLREYNLRPLIKEPAPGKPFPGGADHNINLNFAFDPATVLFTANNYTFVPPT
|         |         |         |         |         |         |         |
 280:      290:      300:      310:      320:      330:      340:

  350:A     360:A     370:A     380:A     390:A     400:A     410:A
  |         |         |         |         |         |         |
VPVLLQILSGAQTAQDLLPAGSVYPLPAHSTIEITLPATALAPGAPHPFHLHGHAFAVVRSAGSTTYNYN
|||||||||...|.||.||||:|.:.....:|:|:||...|  .|||.|||||.|||||||.|||||.
VPVLLQILSGTRDAHDLAPAGSIYDIKLGDVVEVTMPALVFA--GPHPMHLHGHSFAVVRSAGSSTYNYE
|         |         |         |         |         |         |         |
 350:      360:      370:      380:      390:      400:      410:

  420:A     430:A     440:A     450:A     460:A     470:A     480:A
  |         |         |         |         |         |         |
DPIFRDVVSTGTPAAGDNVTIRFQTDNPGPWFLHCHIDFHLEAGFAIVFAEDVADVKAANPVPKAWSDLC
.|:.||||.|... .|||||||..||.||||||||||.||:.|||:|||||.|.....|||||.||.|||
NPVRRDVVSIGDDP-TDNVTIRFVADNAGPWFLHCHIDWHLDLGFAVVFAEGVNQTAVANPVPEAWNDLC
  |         |         |         |         |         |         |
  420:      430:      440:      450:      460:      470:      480:

  490:A
  |
PIYDGLSEAN
|||.......
PIYNSSNPSK
   |
   490:
```

Figure 6: Sequence alignment resulting from structural superimposition of model 3 with the template structure. | means alignment of identical amino acids, : of similar amino acids.

To obtain the functional annotation of the target protein it is necessary to check if in the model the main biological properties of the template are kept. For this reason, with RasMol I analyse the residues in the active site (histidine and cysteine which coordinate the copper), the cysteine involved in the disulphide bonds and the glycosylation pattern. The results are summarized in Table 4. It is possible to assert that the residues in the active site are conserved because the 9 histidine and the cysteine which coordinate the coppers are still present and their distance from the coppers are very similar to those in the template (Table 1). The 4 cysteine involved in the 2 disulphide bonds are also conserved then these structural elements are conserved yet. There is a difference in the glycosylation pattern because only 3 asparagines in the model kept the bond

with NAG. This difference can be related to the different expression systems in which the two proteins are expressed.

Residues	Position in the model	Info	Distance (Å)
his	64	Cu1	2,00
his	66	Cu2	2,15
his	109	Cu2	2,13
his	111	Cu3	2,25
his	394	Cu4	2,02
his	397	Cu1	1,96
his	399	Cu3	2,12
his	450	Cu3	2,16
cys	451	Cu4	2,18
his	452	Cu2	2,19
his	456	Cu4	2,04
cys	85	DB	2,02
cys	486		
cys	117	DB	2,02
cys	204		
asn		Glicosilation	
asn		Glicosilation	
asn		Glicosilation	
asn		Glicosilation	
asn	250	Glicosilation	1,43
asn		Glicosilation	
asn	342	Glicosilation	1,52
asn	434	Glicosilation	1,45

Table 4: In this table are summed up all the information about the most important residues in the model 3. The distances are obtained using RasMol. Comparing it with the one with the data from the template (Table1) I can assert that both the positions and the distances are very conserved.

Then it is possible to assert that the model kept the elements that allow the biological activity in the template and it is necessary to make a comparison between the spatial distribution of them

With the structural alignment obtained from jCE it is possible to superimpose the two structures (model and template) and then to analyse the residues in the active site and in the disulphide bond (Figure 7). It is observed that in these regions the most important amino acids are well conserved with a similar distribution in the space that allows the maintenance of biological activity.

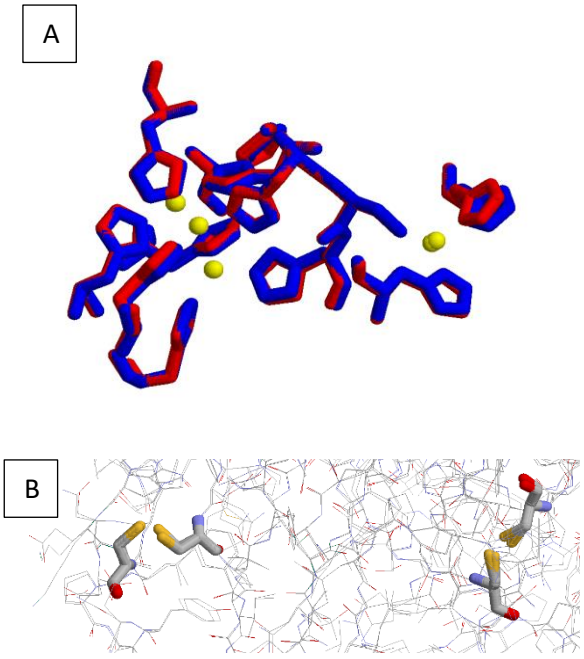

Figure 7: **A**. Superimposition between the 9 histidine and 1 cysteine in the active site from both the template and the model. **B**. Superimposition between the 4 cysteine involved in the disulphide bonds from both the two structures.
Both these images are obtained using RasMol

Discussion

The aim of this work is to functionally annotate a biological entity (Laccase from *Pleurotus eryngii*), through comparison of the structure from another entity with high sequence identity. At the end the goal is to establish whether it is legitimate to transfer GO terms from the template to the target. The generated model for the target Laccase shows overall high structural similarity with the template: they have very high sequence identity and positional conservation of the most structurally and biologically relevant residues. The model retains the four copper ions with the same geometry of coordination and all the histidine and cysteine have similar position. Thus, it is possible to infer that the laccase expressed by *P. eryngii* (the target) is almost equal in function to that of *T. versicolor* (the template). For this reason, it is possible to transfer the GO terms related to molecular function and biological process.

Furthermore, also the cysteine involved in the disulphide bonds are conserved both in the sequence and in the spatial geometry. This observation allows to suppose that not only the conformation of the active site is conserved but the structure of the whole protein does not change. Knowing that the main structural

elements kept their characteristics, also the GO terms related to cellular component can be transferred.

In conclusion, knowing that both species are fungi and they are involved in the same biological processes, I can assert that all the GO terms assigned to the *T. versicolor* laccase can be transferred to the *P. eryngii* laccase:

- Molecular function
  - Copper ion binding (GO:0005507)
  - Hydroquinone: oxygen oxidoreductase activity (GO:0052716)
- Biological process
  - Lignin catabolic process (GO:0046274)
- Cellular component
  - Extracellular region (GO:0005576)

References

1. https://www.uniprot.org/
2. https://www.uniprot.org/help/protein_existence
3. https://enzyme.expasy.org/cgi-bin/enzyme/enzyme-search-ec
4. Piontek K., Antorini M. & Choinowski T. Crystal Structure of a Laccase from the Fungus Trametes versicolor at 1.90-Å Resolution Containing a Full Complement of Coppers. *J. Biol. Chem*. 277, 37663–37669 (2002).
5. https://www.rcsb.org/
6. https://www.uniprot.org/blast/
7. Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25, 3389–3402 (1997).
8. https://embnet.vital-it.ch/software/LALIGN_form.html
9. https://embnet.vital-it.ch/software/LALIGN.man.html
10. https://salilab.org/modeller/
11. Sˇali, A. A Program for Protein Structure Modeling Release 9.23, r11522. 288.
12. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology* 234, 779–815 (1993).
13. http://source.rcsb.org/docu.jsp
14. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering Design and Selection* 11, 739–747 (1998).
15. https://servicesn.mbi.ucla.edu/PROCHECK/
16. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* 26, 283–291 (1993).
17. http://www.rasmol.org/software/RasMol_Latest_Manual.html
18. BLAST: improvements for better sequence analysis | *Nucleic Acids Research* | Oxford Academic. https://academic.oup.com/nar/article/34/suppl_2/W6/2505541.
19. Golding, B., Morton, D. & Haerty, W. *Elementary Sequence Analysis*. 32 (2013).
20. Mount, D. W. Using BLOSUM in Sequence Alignments. *Cold Spring Harb Protoc* 2008, pdb.top39 (2008).
21. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng Des Sel* 12, 85–94 (1999).
22. https://salilab.org/modeller/9v8/manual/node454.html

Supplementary

Figure S1: Sequence global alignment resulting from LALIGN (the character ":" indicates an alignment between identical residues and "." indicates an alignment between similar residues). It is visible the presence in the UniProt sequence of a signal peptide in the beginning where there is no pairing. Despite this the two sequences have an identity of 56.8% which means an high similarity and it allows to use the template for the functional annotation of the target.