3D modelling and functional annotation of the cellular tumor antigen p53 from Bos taurus

Matteo Bolner

February 4, 2019

Abstract

In this project, I would like to verify whether the annotation of P67939 as a p53 cellular tumor antigen can be supported by a more thoroughful search, according to the comparison with a given template of the p53 family. In order to do this, I adopted a procedure in which functional annotation is done after the evaluation of the structure of the protein on the basis of a given template of the family. The procedure was initiated by adopting homology by comparison. With this method, I was able to demonstrate that P67939 is indeed a p53 protein; my results indicate that the protein can be endowed with some of the GO terms associated to the template.

1 Introduction

P67939 is a gene that codifies for a protein of the mammal *Bos taurus* which, after manual annotation of the transcript, is believed to be a member of the p53 cellular tumor antigen family. P53 proteins are transcription factors implicated in the regulation of the cell cycle, particularly through the monitoring of DNA integrity before replication. Their function is thought to be tumour suppression, which would justify why in many types of cancer p53 is mutated or inactivated[1].
The structure of p53 is organized in four different domains:
1)a N-terminal transactivation domain, which allows the binding of other proteins such as transcription coregulators;
2)a DNA-binding domain, which binds specific DNA sequences;
3)an oligomerization domain, involved in the assembly of p53 monomers into tetrameric structures;
4)a C-terminal regulatory domain.
Additionally, a tightly bound zinc atom necessary for the DNA binding activity is present[2].
According to article [1], the binding of DNA is possible thanks to an induced fit mechanism involving a conformational switch in the loop L1 of the DNA binding domain. The conformational switch happens only in the subunits contacting the inner repeats of the targeted DNA sequence, as is also shown in the reference article [1] In the subunits binding the outer DNA repeats, the L1 loop mantains a recessed conformation.
Each DNA-binding domain interacts with the DNA-binding domain of another subunit through hydrogen bonds. For a more detailed description of the structure of p53 see article [1].

2 Methods

2.1 Data bases

The P67939 target sequence is obtained from the UniProt database (release 2018_11)[3] ; the 3KMD template sequence and structure are obtained from the PDB file in RCSB PDB (18.12.18 release)[4]. The GO terms are obtained from the UniProt page associated to p53. Information in relation to the protein domains and secondary structure is derived from InterPro (v.71, 8.11.18)[5].

2.2 Computational methods

The template protein is found after local alignment of the target (P67939) with the UNIPROTKB_PDB database with BLASTP (2.7.1+)[6], 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 [7] with the parameters highlighted in Table 2.

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

Alignment method	Local
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 [8] and then used to generate five model structures of the target. Pymol is adapted as the molecular visualizer (Version 2.2.0)[9]. The online version of PROCHECK 3.5 [10] analyzes the stereochemical quality of the model. The computed target structure is aligned with the template structure with jCE (V2.11 2014 March 17)[11] installed locally.

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; I then choose P04637, with a sequence identity of 80.2%, a 0.0 E-value and the highest score (1,622). The whole sequence is covered by this preliminary alignment. On PDB, P04637 corresponds to many entries; the 3KMD crystal [12], obtained with a resolution of 2.15 Å and manually annotated, contains all the four subunits comprised in the functional unit of p53 and a DNA sequence bound to the monomers; it also has no engineered mutations that might change its functionality. The structure covers 51% of the protein sequence, and the tetramerization domain is not present in the crystal. However, reading through its reference article I determine that its presence is not necessary, since the four monomers are self-assembled into a tetramer. I therefore choose 3KMD as the best suiting template to represent the p53 protein in this project.

4 Sequence alignment

Figure 1 shows the alignment of the template with the target sequence. The The sequence identity is 90.5% with a sequence alignment length of 201 residues, and a coverage of 70.4%. Almost all of the important residues described in the introduction are conserved, as is shown in the figure; in particular, all the critical residues of the DNA-binding domain are perfectly conserved.

```
>>P67939 386 bp                                                (386 aa)
 Waterman-Eggert score: 1268;  310.1 bits; E(1) <  3.4e-89
90.5% identity (95.5% similar) in 201 aa overlap (1-200:84-284)

               10        20        30        40        50        60
3KMD   PLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTP
       :::: ::::::: :.:::::::::.::::::::::::.:::.:::::::::::::::: :
P67939 PLSSFVPSQKTYPGNYGFRLGFLQSGTAKSVTCTYSPSLNKLFCQLAKTCPVQLWVDSPP
            90        100       110       120       130       140

               70        80        90        100       110
3KMD   PPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSD-SDGLAPPQHLIRVEGNLRVEYLDDRN
       :::::::::::::. .::::::::::::::: :: ::::::::::::::::::.::::::
P67939 PPGTRVRAMAIYKKLEHMTEVVRRCPHHERSSDYSDGLAPPQHLIRVEGNLRAEYLDDRN
           150       160       170       180       190       200

     120       130       140       150       160       170
3KMD   TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSF
       :::::::::::: :::. :.::::::::.:::::::::::::::::::::::: :::::::::
P67939 TFRHSVVVPYESPEIDSECTTIHYNFMCNSSCMGGMNRRPILTIITLEDSCGNLLGRNSF
           210       220       230       240       250       260

     180       190       200
3KMD   EVRVCACPGRDRRTEEENLRK
       :::::::::::::::::::::
P67939 EVRVCACPGRDRRTEEENLRK
           270       280
```

Figure 1: Alignment of the template and target sequences with Lalign; the conserved residues of the DNA-binding domain are highlighted in green, while the rest of the conserved residues are highlighted in yellow. Not conserved residues are highlighted in red.

5 Modeller at work

5.1 Input preparation

The global alignment derived from Lalign was converted into .pir format by following Modeller's instructions. The number of models to produce was set to 5 in the Modeller script.

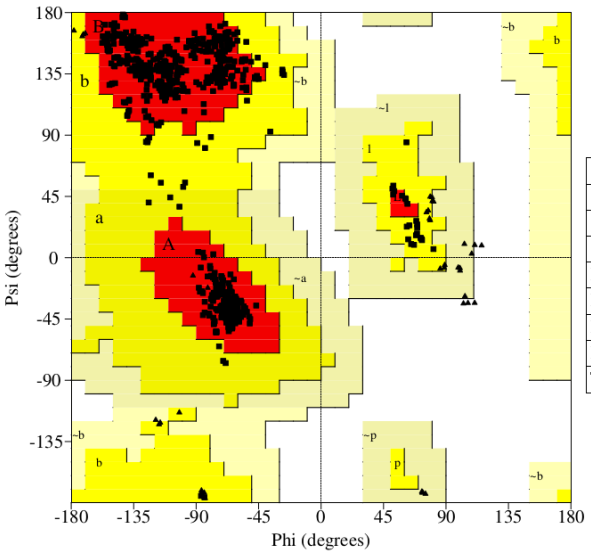
5.2 Modeller output

Table 3 lists the scores of the five different models obtained: molpdf (molecular probability density function), DOPE (discrete optimized protein energy) and GA341 (score that uses the percentage sequence identity between the template and the model as a parameter) . The scores are computed according to Modeller's internal validation procedure. TARGET.B99990003, having the lowest molpdf score and the highest DOPE score is adopted as the final result for modelling the structure of the target.

Table 3: Modeller output table

Filename	molpdf	DOPE score	GA341 score
TARGET.B99990001.pdb	6519.65674	-79662.54688	1.00000
TARGET.B99990002.pdb	6377.20801	-79359.34375	1.00000
TARGET.B99990003.pdb	6334.77441	-79956.13281	1.00000
TARGET.B99990004.pdb	6752.84521	-79290.57031	1.00000
TARGET.B99990005.pdb	6596.75293	-79066.58594	1.00000

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 2) the model is stable, with 92.3% of the residues having the proper torsion angles and being accordingly in the most favoured regions of the main chain backbone.



Residues in most favoured regions [A,B,L]	635	92.3%
Residues in additional allowed regions [a,b,l,p]	53	7.7%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	688	100.0%
Number of end-residues (excl. Gly and Pro)	46	
Number of glycine residues (shown as triangles)	48	
Number of proline residues	64	
Total number of residues	846	

Figure 2: Ramachandran plot and statistics of the model obtained

6 Target annotation and discussion

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

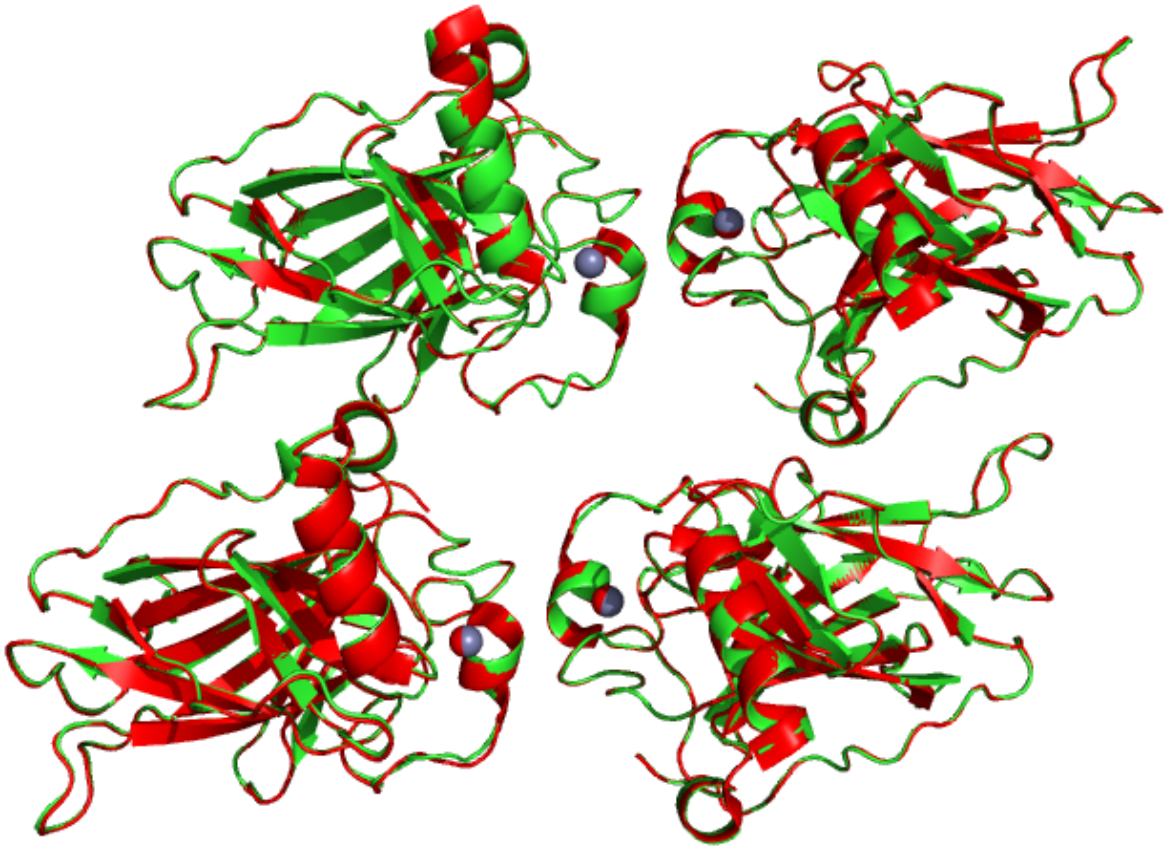


Figure 3: Superimposition of template(red) and model structure(green)

Figure 3 shows the superimposition of the template structure with the target model; the figure indicates that secondary structural motives are conserved. The RMSD is 0.22 Å.

```
Twists 0 ini-len 800 ini-rmsd 0.25 opt-equ 800 opt-rmsd 0.22 chain-rmsd 0.22 Score 2243.73 align-len 804 gaps 4 (0.50%)
Z-score 8.57 Afp-num 5 Identity 90.55% Similarity 94.53%
Block  0 afp  0 score  0.00 rmsd  0.22 gap 0 (NaN%)

                .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:   92 PLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMA
               ||||.|||||||.|.|||||||.|||||||||||||.|||:||||||||||||||.||||||||||
Chain 2:   92 PLSSFVPSQKTYPGNYGFRLGFLQSGTAKSVTCTYSPSLNKLFCQLAKTCPVQLWVDSPPPPGTRVRAMA

                .    :    .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:  162 IYKQSQHMTEVVRRCPHHERCSD-SDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCT
               |||..:|||||||||||||.|| ||||||||||||||||.|||||||||||||||||.||:.|:||
Chain 2:  162 IYKKLEHMTEVVRRCPHHERSSDYSDGLAPPQHLIRVEGNLRAEYLDDRNTFRHSVVVPYESPEIDSECT

                .    :    .    :    .    :    .    :    .    :    .    :
Chain 1:  231 TIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRK
               |||||:|||||||||||||||||||||.||||||||||||||||||||||||||||||
Chain 2:  232 TIHYNFMCNSSCMGGMNRRPILTIITLEDSCGNLLGRNSFEVRVCACPGRDRRTEEENLRK
```

Figure 4: Sequence alignment of template (chain 1) and model (chain 2) derived from structural superimposition by jCE

Figure 4 shows the sequence alignment derived with jCE from the structural superimposition. The structure of the template compares very well with the structure of the model: the sequence

identity as derived after structural alignment is 90.5%. In order to verify whether the important residues described in the reference article[14] are conserved, the superimposition is zoomed to the point to which they are visible and comparable.

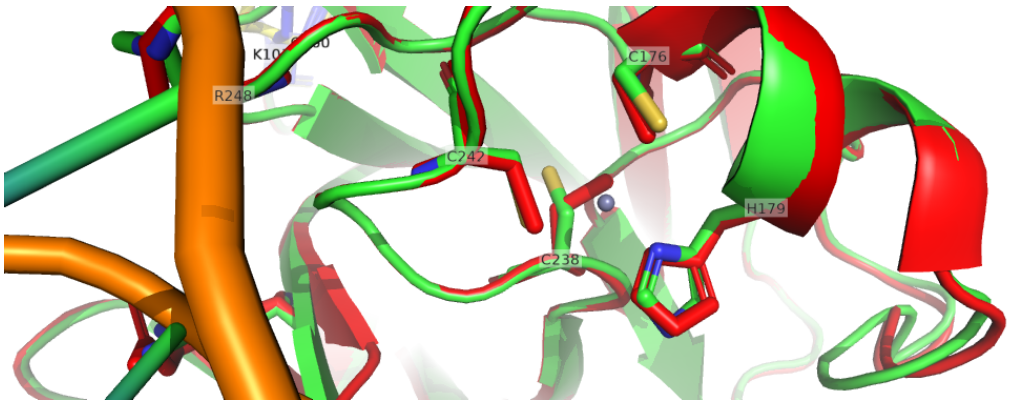


Figure 5: Zinc-binding residues of template(red) and model structure(green)

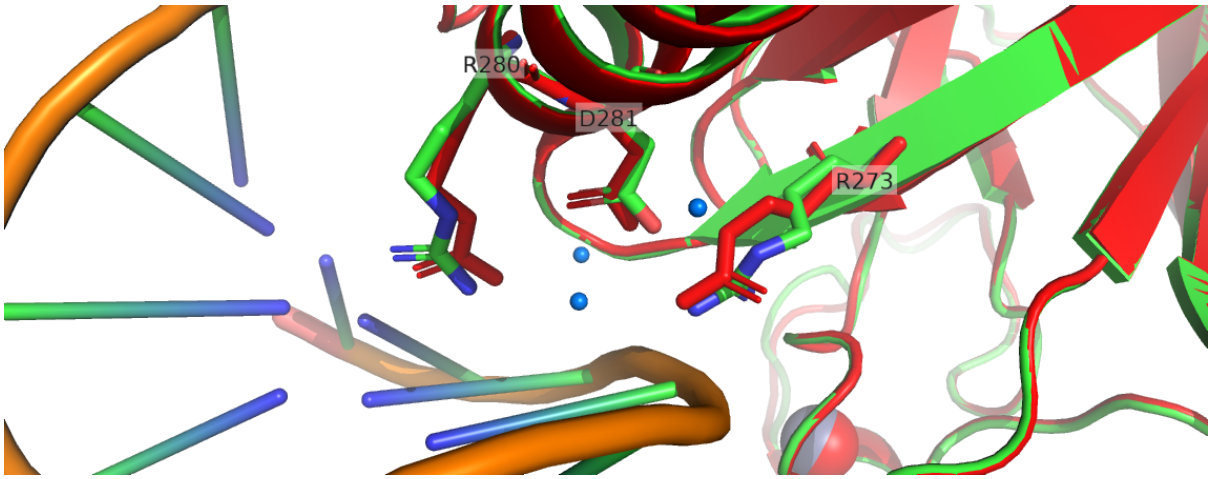


Figure 6: Residues of beta-sheet 10 and alpha-helix 1 interacting with water molecules(blue spheres) and DNA. The template is red and the model is green.

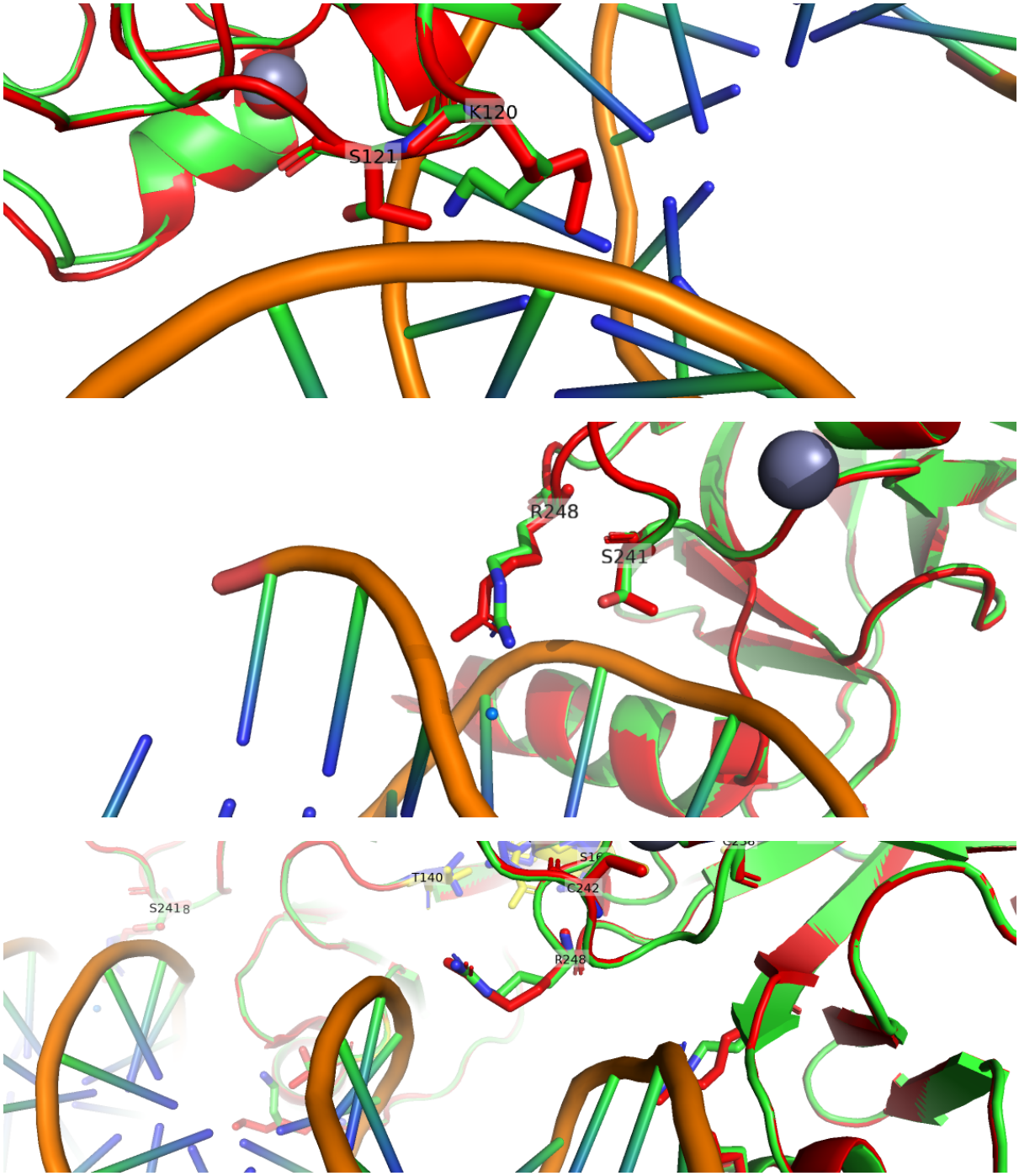


Figure 7: Residues involved in the binding of DNA; the second subfigure belongs to chains B and D, while the third belongs to chains A and C. The difference in conformation between the different chains is conserved.

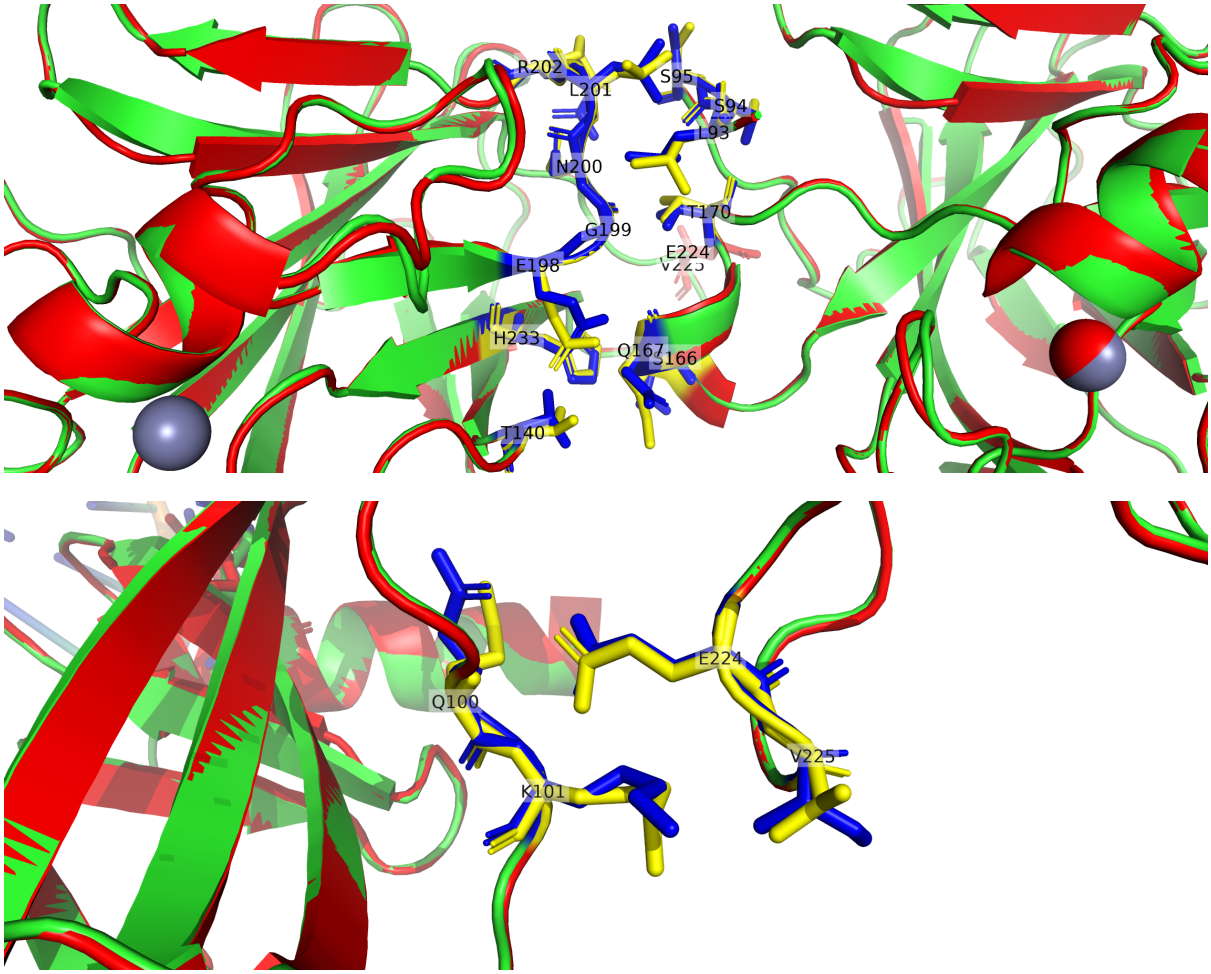


Figure 8: Residues involved in the interaction between monomers of the protein; the yellow residues belong to the template(red), while the blue ones belong to the model(green

I can confirm that almost all important residues are conserved; therefore I may conclude that the function of the DNA-binding domain is conserved. Additionally, the zinc atom position is conserved, along with the residues bound to it.

The tetramerization domain is not crystallized in 3KMD; however, I had previously produced a model using 3Q05 as template structure, which contains the tetramerization domain. The model was deemed unacceptable because of the many engineered mutations that might have impaired its functionality; the tetramerization domain is very well conserved, as seen in figure 9, even if two engineered mutations were inserted. This allows me to predict that, while the tetramerization domain is missing from the final model, it is still conserved in the target protein, since the sequence of 3Q05 is the same as 3KMD.

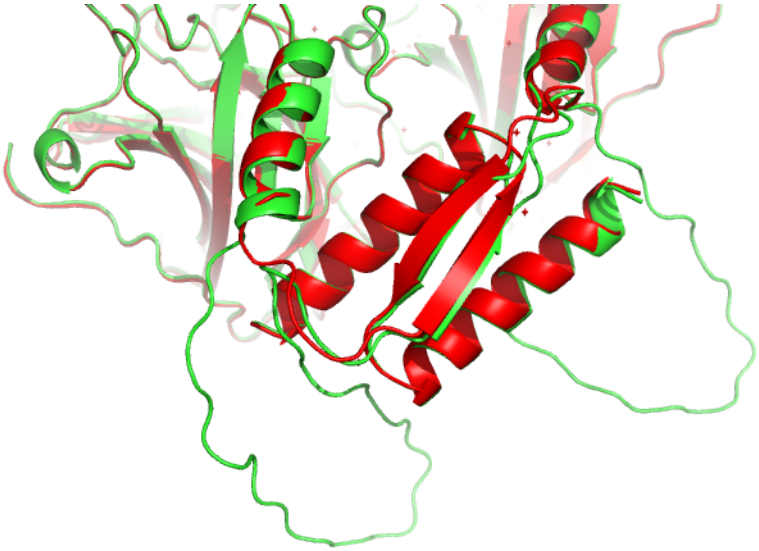


Figure 9: Tetramerization domains of 3Q05(red) and model(green)

Having confirmed the conservation of the important residues, 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, after the computation of a model structure,I have all the necessary information to prove that the protein can function as a p53. Considering the GO terms associated to the DNA-binding domain of p53, I may transfer them to my target.

Table 4: GO terms associated to p53 whose conservation was demonstrated in the target model.

Molecular function
DNA binding
DNA-binding transcription factor activity
DNA-binding transcription factor activity, RNA polymerase II-specific
DNA-binding transcription activator activity, RNA polymerase II-specific
Disordered domain specific binding
Zinc ion binding
Identical protein binding
Protein heterodimerization activity

7 Swiss-model

In order to verify whether manual annotation of the target sequence can be replaced by automatic methods, I submit the target to the Swiss-model[13]; however, its first choice of template is 3Q05, which was initially rejected. I can therefore conclude that in the case of this protein, automatic methods are still not perfected enough to replace the manual procedure.

8 References

1 An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. Petty, T.J., Emamzadah, S., Costantino, L., Petkova, I., Stavridi, E.S., Saven, J.G., Vauthey, E., Halazonetis, T.D. (2011) Embo J. 30: 2167-2176

2 Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Cho Y1, Gorina S, Jeffrey PD, Pavletich NP. Science. 1994 Jul 15;265(5170):346-55.

3 https://www.uniprot.org/uniprot/P67939

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

4 https://www.rcsb.org

- "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

5 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

6 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

7 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.

8 https://salilab.org/modeller/

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

9 https://pymol.org

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

10 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]

11 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

12 https://www.rcsb.org/structure/3q05

12 https://www.swissmodel.expasy.org

- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. "SWISS-MODEL: homology modelling of protein structures and complexes." Nucleic Acids Res. 46(W1), W296-W303 (2018).

14 Chen, Y., Dey, R., Chen, L. "Crystal structure of the p53 core domain bound to a full consensus site as a self-assembled tetramer." (2010) Structure 18: 246-256