Analysis of a Protein-Ligand Complex (1c14)

Ioannis Valasakis[1]

[1]Birkbeck, University of London

March 10, 2019

Plagiarism statement

I, Ioannis Valasakis, have read the advice sections on plagiarism course web site and I confirm that I understand them. I also confirm that the work contained within this essay is my own effort and I fully acknowledge and reference the work of others within this essay.

Dated 16/02/2019

Please note:-

Plagiarism: Coursework and projects submitted by candidates must be expressed in their own words and incorporate their own ideas and judgments. No large scale copying of sections of text is acceptable, with or without proper acknowledgement. Students must convey ideas in their own words, only copying short sections of text if the exact wording is important. Any copying must be clearly indicated by the use of quotation marks and a literature reference. The presentation of another person's thoughts or words or software or diagrams or pictures as if it were the candidates own is plagiarism and will incur a severe academic penalty in the assessment of the submitted work.

Protein 1c14

Part A

The structure of reference for this coursework is of the protein with PDB code '1c14' namely the Enoyl-ACP from Escherichia coli reductase-NAD+-triclosan complex and its structural classification is oxidoreductase. There are various kinds of oxidoreductases including peroxidases, hydroxylases, oxygenases, and reductases. Peroxidases are localized in peroxisomes, and they catalyze the reduction of hydrogen peroxide. Hydroxylases add hydroxyl groups to its substrates. Oxygenases incorporate oxygen from molecular oxygen into organic substrates. Reductases catalyze reductions, and in most cases, reductases can act as an oxidase. See Fig 1.

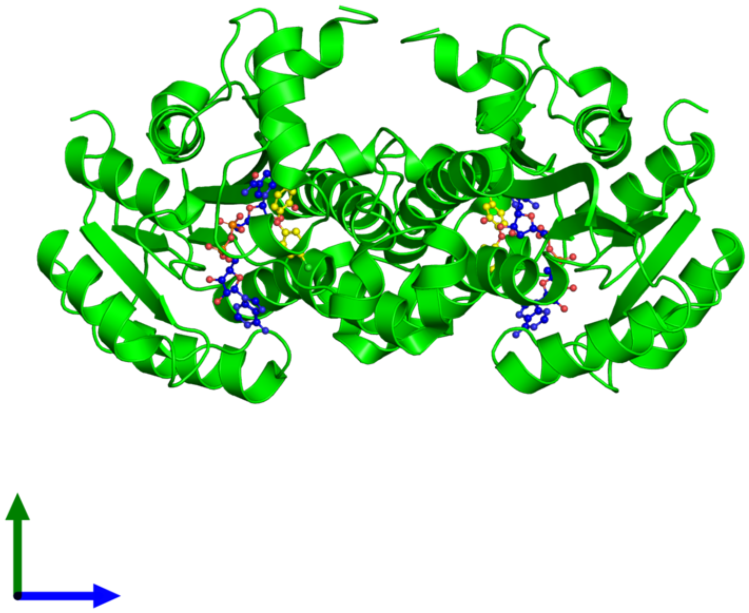


Figure 1: Protein 1c14

Question a

Oxidoreductases specifically catalyze the transfer of electrons from one molecule (the oxidant) to another molecule (the reductant). Oxidoreductases catalyze reactions similar to this example: A– + B A + B– where A is the oxidant and B is the reductant. Oxidoreductases can be oxidases or dehydrogenases.

Here is the equation of the catalysed reaction:

An acyl-[acyl-carrier protein] + NAD(+) = a trans-2,3-dehydroacyl-[acyl-carrier protein] + NADH

Oxidases are enzymes involved when molecular oxygen acts as an acceptor of hydrogen or electrons. On the other hand, dehydrogenases are enzymes that oxidize a substrate by transferring hydrogen to an acceptor that is either NAD+/NADP+ or a flavin enzyme.

Question b

The structural classification of the protein is oxidoreductase. The lineage viewed from *the Structural Classification of Proteins* (SCOP) (Conte, 2000) database is as follows:

- Root: scop [the main database]
  - Class: Alpha and beta proteins (a/b)
    - Fold: NAD(P)-binding Rossmann-fold domains

- Superfamily: NAD(P)-binding Rossmann-fold domains
    - Family: Tyrosine-dependent oxidoreductases
        - Protein: Enoyl-ACP reductase
            - Species: Escherichia coli

Question c

Here is the generated secondary 1c14 structure chain and 256 residues shown in Fig 2.

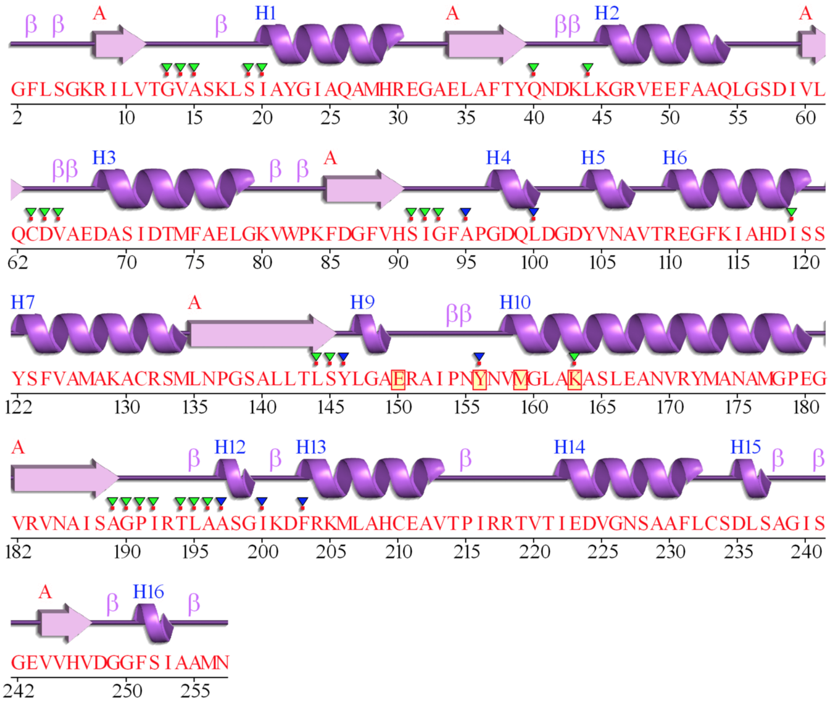

β β A β H1 A ββ H2 A
GFLSGKRILVTGVASKLSIAYGIAQAMHREGAELAFTYQNDKLKGRVEEFAAQLGSDIVL
2 10 15 20 25 30 35 40 45 50 55 60
ββ H3 β β A H4 H5 H6
QCDVAEDASIDTMFAELGKVWPKFDGFVHSIGFAPGDQLDGDYVNAVTREGFKIAHDISS
62 70 75 80 85 90 95 100 105 110 115 120
H7 A H9 ββ H10
YSFVAMAKACRSMLNPGSALLTLSYLGAERAIPNYNVMGLAKASLEANVRYMANAMGPEG
122 130 135 140 145 150 155 160 165 170 175 180
A β H12 β H13 β H14 H15 β β
VRVNAISAGPIRTLAASGIKDFRKMLAHCEAVTPIRRTVTIEDVGNSAAFLCSDLSAGIS
182 190 195 200 205 210 215 220 225 230 235 240
A β H16 β
GEVVHVDGGFSIAAMN
242 250 255

Figure 2: Chain and 256 residues

From the ProMotif (Dasgupta *et al.*, 2007) database it can be observed the following:

- 1 beta sheet
- 5 beta-alpha-beta motifs
- 1 beta bulge
- 7 strands
- 16 helices
- 22 helix-helix interaction, and
- 18 beta turns.

In Fig 3 there is also depicted the Hera diagram of main chain H-bonds

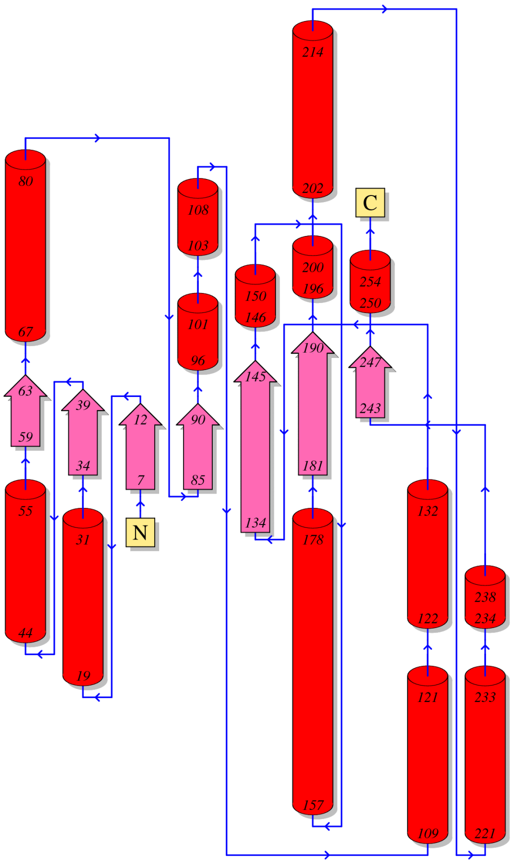


Figure 3: domain diagram of 1c14

Question d

Is the structure of good quality? Three different ways have been explored using the ProSA website in order to calculate an overall quality score for the specific input structure.

Z-Score

The z-score indicates overall model quality with its value displayed in a plot that contains the z-scores of all experimentally determined protein chains in current PDB. There are groups of structures from different sources (X-ray, NMR) which are distinguished by colouring. The Z-score can be used to check whether it is within the range of scores typically found for native proteins of similar size, which seems to be in our case (see Fig 4) with Z = -8.97

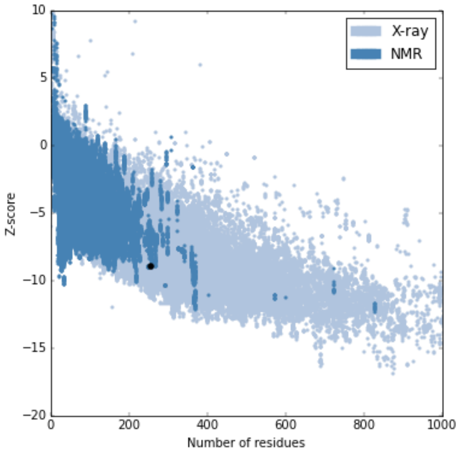


Figure 4: Z-score

Plot of residue scores

Here is depicted the local model quality by plotting energies as a function of amino acid sequence position *i*. Positive values correspond to erroneous or problematic parts of the structure. A plot of single residue energies usually contains large fluctuations and is of limited value for model evaluation. Therefore the plot is smoothed by calculating the average energy over each 40-residue fragment *s(i,i+39)*, which is then assigned to the 'central' residue of the fragment at position *i+19.* A second line with a smaller window size of 10 residues is shown in the background of the plot.

Interactive molecule viewer

ProSA-web visualizes (Bhargavi *et al.*, 2017) the 3D structure of the input protein using the molecule viewer Jmol. Residues are coloured from blue to red in the order of increasing residue energy.

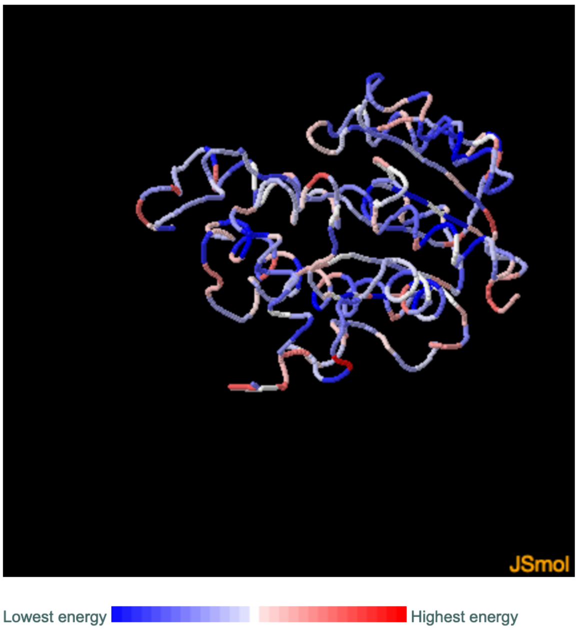


Figure 5: Interactive energy related mol viewer

Residue-property plots

These plots are drawn for all protein, RNA and DNA chains in the entry. The first graphic for a chain summarises the proportions of the various outlier classes displayed in the second graphic. The second graphic shows the sequence view annotated by issues in geometryand electron density. Residues are color-coded according to the number of geometric quality criteria for which they contain at least one outlier: green = 0, yellow = 1, orange = 2 and red = 3 or more. A red dot above a residue indicates a poor fit to the electron density (RSRZ > 2). Stretches of 2 or more consecutive residues without any outlier are shown as a green connector. Residues present in the sample, but not in the model, are shown in grey.

Ramachandran Plot

The statistics from the Ramachandran Plot depicted in Fig. 6 are shown on the table below:

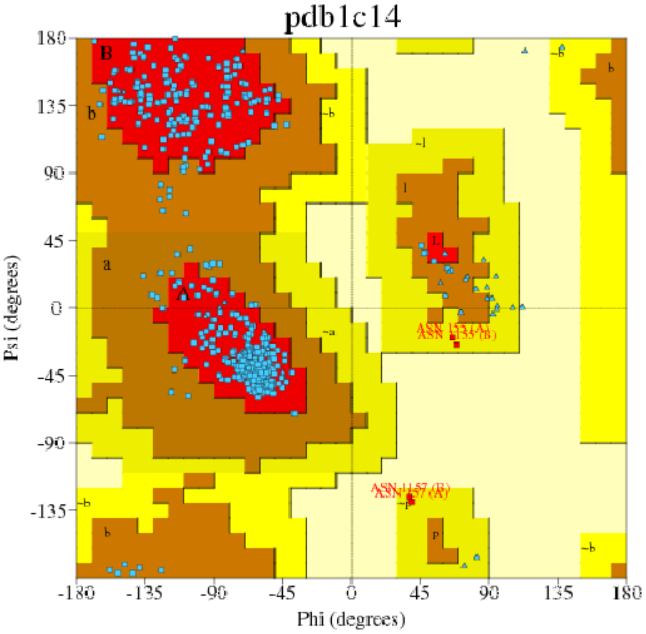


Figure 6: Ramachandran Plot of 1c14

```
                                              No. of
                                             residues      %-tage
                                              ------       ------
Most favoured regions       [A,B,L]             397         89.0%*
Additional allowed regions  [a,b,l,p]            45         10.1%
Generously allowed regions  [~a,~b,~l,~p]         4          0.9%
Disallowed regions          [XX]                  0          0.0%
                                               ----        ------
Non-glycine and non-proline residues            446        100.0%

End-residues (excl. Gly and Pro)                  2

Glycine residues                                 50
Proline residues                                 14
                                               ----
Total number of residues                        512
```

From the PROCHECK statistics and their analysis of 118 structures of resolution of at least 2.0 Angstroms and *R*-factor no greater than 20.0 a good quality model would be expected to have over 90% in the most favoured region.

G-factors

G-factors provide a measure of how unusual, or out-of-the-ordinary, a property is. Specifically, values below -0.5* denote unusual properties while if they are below -1.0** they are considered highly unusual.

```
Parameter                         Score      Average
                                             Score
---------                         -----      -----
Dihedral angles:-
     Phi-psi distribution         -0.11
     Chi1-chi2 distribution       -0.23
     Chi1 only                    -0.05
     Chi3 & chi4                   0.44
     Omega                         0.47
                                              0.12
                                             =====
Main-chain covalent forces:-
     Main-chain bond lengths       0.57
     Main-chain bond angles        0.32
                                              0.42
                                             =====

     OVERALL AVERAGE                          0.25
                                             =====
```

In the specific case, the overall average is 0.25 which is much higher than the threshold for an unusual property value, which is another indication of the quality of the model.

Therefore from all the above sources, we can confidently conclude that the specific model has a quite good quality.

Question e

The co-factor is Nicotinamide-Adenine-Dinucleotide (NAD) and the inhibitor-like ligand is triclosan, which is also depicted in Fig 8 and viewed from the top in Fig 7.

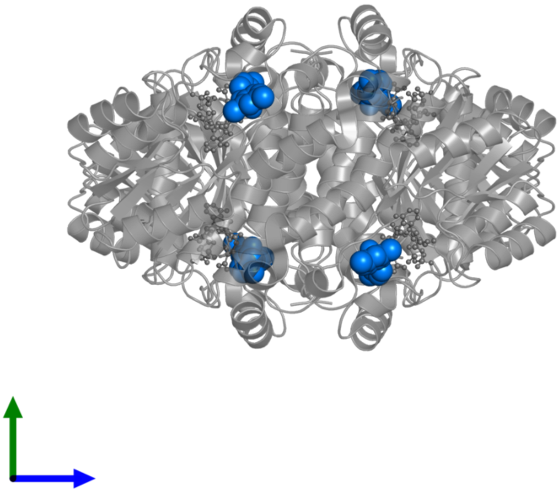


Figure 7: Triclosan viewed from the top (3D).

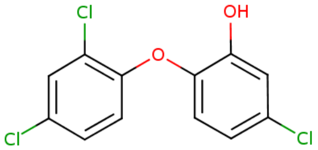


Figure 8: Triclosan

Question f

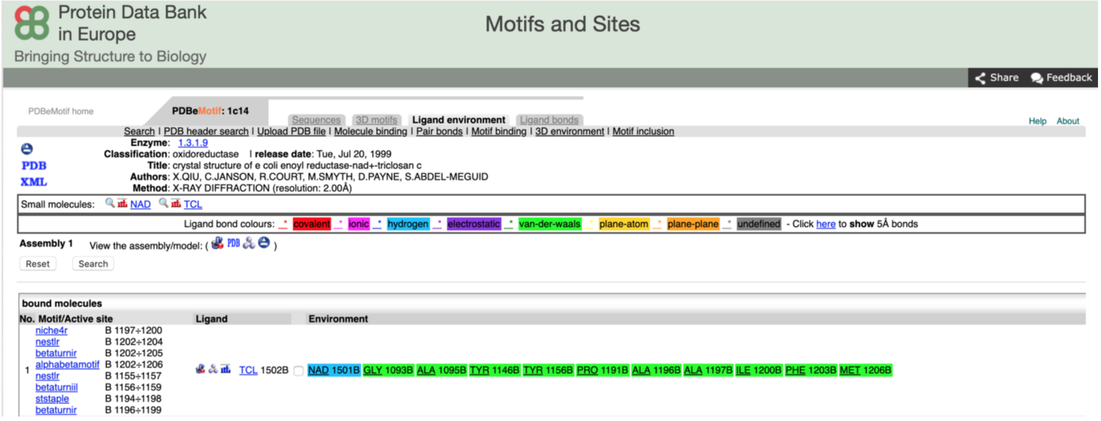
In this part the EMBL/EBI Motif Tool (Bateman *et al.*, 2002) is used to fetch by ID (*1c14*) the protein and create the biological oligomeric state in order to identify which functional protein groups interact with triclosan and type of their interaction.

In Fig 9 there is the description of the formed Van Der Waals contacts with the depicted residues. Specifically, those are:

- GLY 1093B
- ALA 1095B
- TYR 1146B
- TYR 1156B
- PRO 1191B
- ALA 1196B
- ALA 1197B
- ILE 1200B

- PHE 1203B
- MET 1206B

with a hydrogen bond as expected with the NAD 1501B.



Protein Data Bank in Europe
Bringing Structure to Biology

Motifs and Sites

PDBeMotif: 1c14

Ligand environment

Search | PDB header search | Upload PDB file | Molecule binding | Pair bonds | Motif binding | 3D environment | Motif inclusion

Enzyme: 1.3.1.9
Classification: oxidoreductase | **release date**: Tue, Jul 20, 1999
Title: crystal structure of e coli enoyl reductase-nad+-triclosan c
Authors: X.QIU, C.JANSON, R.COURT, M.SMYTH, D.PAYNE, S.ABDEL-MEGUID
Method: X-RAY DIFFRACTION (resolution: 2.00Å)

Small molecules: NAD TCL

Ligand bond colours: covalent ionic hydrogen electrostatic van-der-waals plane-atom plane-plane undefined - Click here to **show** 5Å bonds

Assembly 1 View the assembly/model: (PDB)

bound molecules

No.	Motif/Active site		Ligand	Environment
1	niche4r	B 1197÷1200	TCL 1502B ☐	NAD 1501B GLY 1093B ALA 1095B TYR 1146B TYR 1156B PRO 1191B ALA 1196B ALA 1197B ILE 1200B PHE 1203B MET 1206B
	nestlr	B 1202÷1204		
	betaturnir	B 1202÷1205		
	alphabetamotif	B 1202÷1206		
	nestlr	B 1155÷1157		
	betaturniil	B 1156÷1159		
	ststaple	B 1194÷1198		
	betaturnir	B 1196÷1199		

Figure 9: Motifs and Sites for TCL in 1c14

Question g

In Fig 10 it’s shown the ligand plot generated using the open source software LeView from Pegase Biosciences.

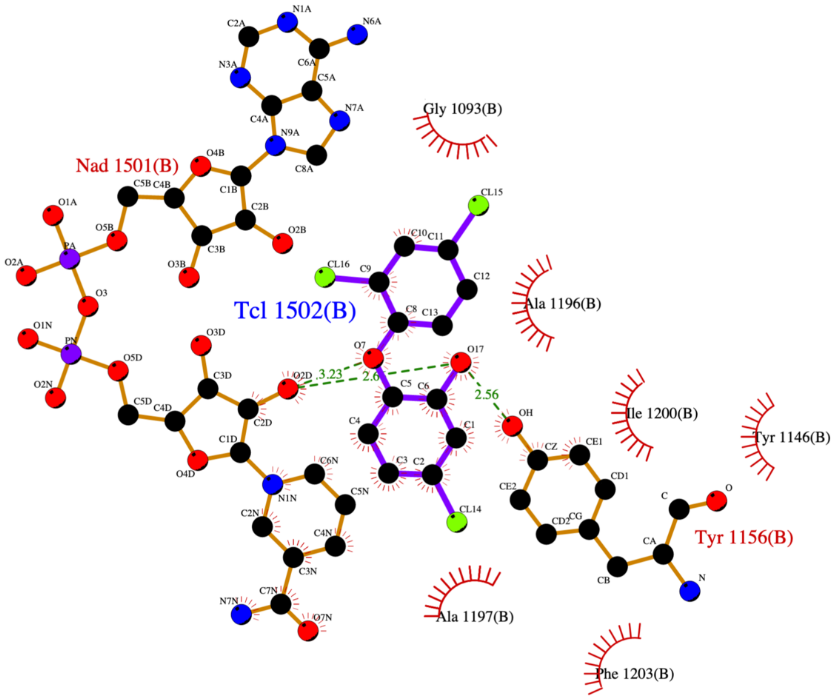


Figure 10: Ligand interaction of A site (TCL)

A lot of ligands are stabilized mostly by hydrogen bonds, but there are also cases with very nonpolar ligands stabilized in the binding site by van der Walls interactions, which seems to be the case in this specific molecule. Therefore, in our specific case, the van der Waals interactions work additively to achieve stability, especially the ones which are closer to the residue (GLY 1093B, ALA 1196B, ILE 1200B, TYR 1146B, ALA 1197B, PHE 1203B).

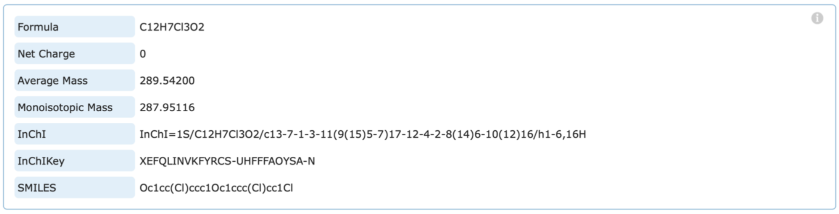
PART B

Question a

Looking up for the triclosan in the Drugbank (www.drugbank.ca) (Wishart *et al.*, 2017)database returns the following SMILES string of the triclosan (TCL):

```
OC1=CC(Cl)=CC=C1OC1=C(Cl)C=C(Cl)C=C1
```

For cross-validation, a search was performed on the EBI CHebi database as shown in Fig 11



Formula	C12H7Cl3O2
Net Charge	0
Average Mass	289.54200
Monoisotopic Mass	287.95116
InChI	InChI=1S/C12H7Cl3O2/c13-7-1-3-11(9(15)5-7)17-12-4-2-8(14)6-10(12)16/h1-6,16H
InChIKey	XEFQLINVKFYRCS-UHFFFAOYSA-N
SMILES	Oc1cc(Cl)ccc1Oc1ccc(Cl)cc1Cl

Figure 11: CHebi results for triclosan

The SMILES result from Question a was saved in a text file, named as ” *TCL.smi*”. This was converted using OpenBabel to an *sdf* file to be used in the filtering later. Additionally, an *sdf* file with 3D optimised coordinates for Triclosan (*TCL_3D.sdf*). The resulting structure was optimized using the given forcefield and checked for the lowest-energy conformer using a Monte Carlo search.

Question b

Using the babel and (Quirós *et al.*, 2018) OpenBabel software installed on a Google Cloud instance of 24 cores and 200GB RAM, running Debian Stretch. It was installed using the Debian *apt* package manager and thus the fingerprinting through the Chembl database was performed. The search on Chembl was filtered to Phenols as shown in the heatmap of Fig. 12

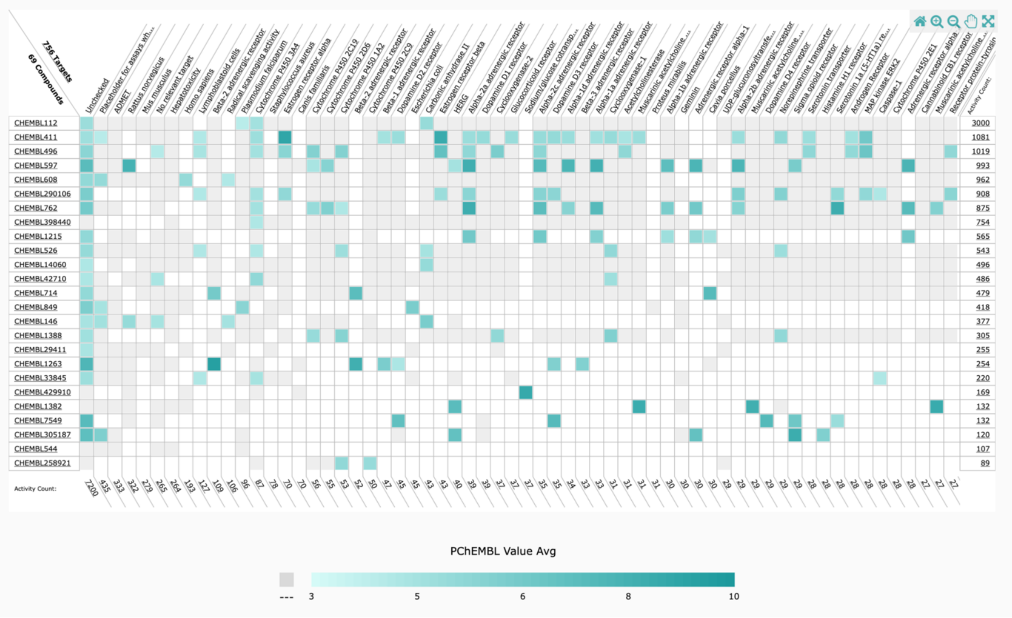


Figure 12: Chembl phenol drug search results

As the subset wasn't compatible with OpenBabel *sdf* parser, the whole drug database was downloaded as *sdf gzip'ped* file and it was randomised by importing it in R (and choosing random lines using a randomised function.

```
create_rsample = function(df,n){
   return(df[sample(nrow(df),n),])
}
```

Question c

Using OpenBabel, the drug database and the SMILES and *sdf* file of Triclosan a database fingerprint performed and the requested values were generated. This was extracted using babel with the command:

```
babel TRC.smi drug_database.sdf -ofpt >> ioannis_app_drugs.txt
```

to a text file (*ioannis_app_drugs.txt).* A Python program was written to extract those data in a comma delimited file that was later imported to R in order to create histograms.

```
import re

def return_regex(txt):
    re='.*?'    # Non-greedy match on filler
    re2='((?:[a-z][a-z]*[0-9]+[a-z0-9]*))'  # Alphanum chars
    re4='([+-]?\\d*\\.\\d+)(?![-+0-9\\.])'  # Float number

    rg = re.compile(re+re2+re+re4,re.IGNORECASE|re.DOTALL)
    m = rg.search(txt)
    if m:
        alphanum1 = m.group(1)
        float1 = m.group(2)

        return (alphanum1 + ',' + float1)

triclosan_dc = 'ioannis_app_drugs.txt'
```

```
ftri = open(triclosan_dc, 'r')
[print(return_regex(line)) for line in ftri]
```

Here is the histogram generated by the R plot:

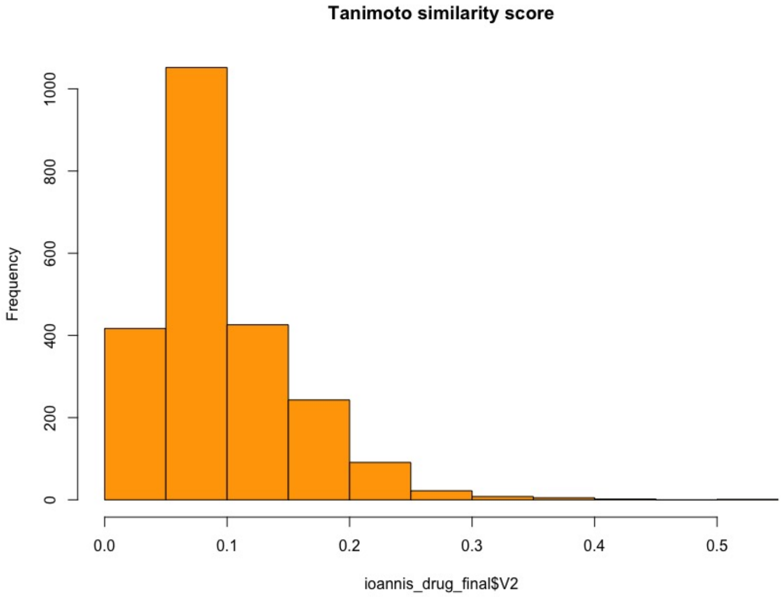


Figure 13: Tanimoto similarity score

Using the *summary* function in R the median = 0.083 and the mean = 0.096 and SD = 0.057

```
> summary(ioannis_drug_final$V2, na.rm=TRUE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.01163 0.05789 0.08264 0.09644 0.12052 0.53226      61
> sd(ioannis_drug_final$V2, na.rm=TRUE)
[1] 0.057
```

In order to find the 5 most similar molecules, OpenBabel was used:

```
iris:data wizofe$ babel approved.fs mostsim.sdf -s triclosan.sdf -at5
5 molecules converted
140 audit log messages
```

To identify their similarity another run of OpenBabel:

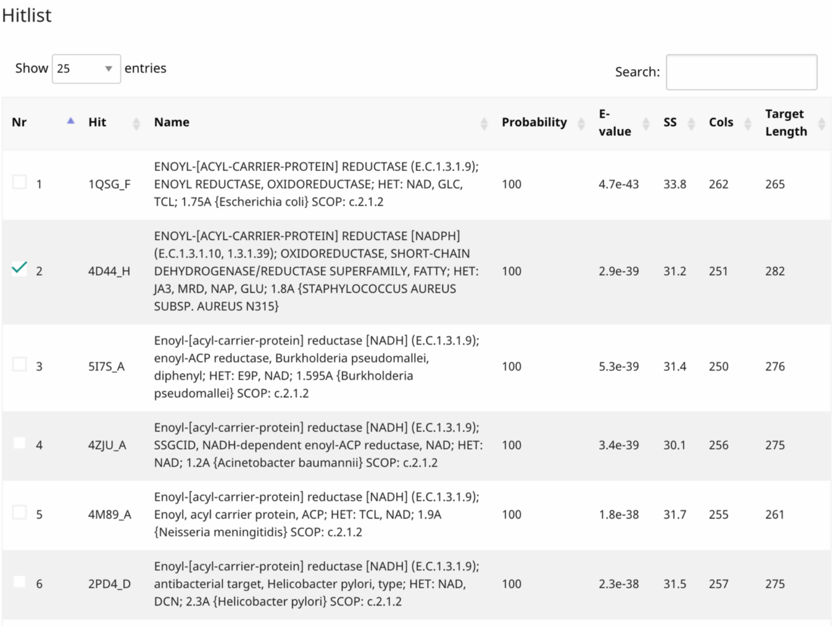
```
iris:data wizofe$ babel triclosan.sdf mostsim.sdf -ofpt
>triclosan
>D00CSQ   Tanimoto from triclosan = 1
Possible superstructure of triclosan
>D06ZAY   Tanimoto from triclosan = 0.532258
>D0J5DC   Tanimoto from triclosan = 0.430769
>D09QDP   Tanimoto from triclosan = 0.428571
>D02VMJ   Tanimoto from triclosan = 0.369565
```

Ignoring the first match (as it is itself, Triclosan) the highest similarity is with the ligand haloprogin (C9H4Cl3IO) with a score of 0.53. This can also be validated with the histogram score above, as it is indeed the maximum value found using the R summary was 0.532, which is the D06ZAY.

PART C

Question a

The best way to identify a homologous structure is to perform an HMM comparison using HHpred of the MBI Bioinformatics Toolkit. The *FASTA* file was downloaded from UniProt and the comparison was performed as shown in Fig 14.



Hitlist

Show 25 entries

Search:

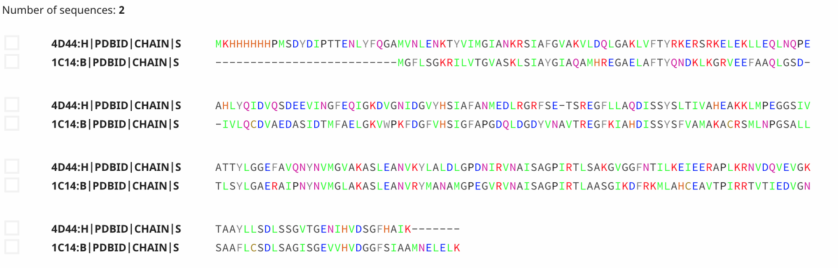
Nr	Hit	Name	Probability	E-value	SS	Cols	Target Length
☐ 1	1QSG_F	ENOYL-[ACYL-CARRIER-PROTEIN] REDUCTASE (E.C.1.3.1.9); ENOYL REDUCTASE, OXIDOREDUCTASE; HET: NAD, GLC, TCL; 1.75A {Escherichia coli} SCOP: c.2.1.2	100	4.7e-43	33.8	262	265
✓ 2	4D44_H	ENOYL-[ACYL-CARRIER-PROTEIN] REDUCTASE [NADPH] (E.C.1.3.1.10, 1.3.1.39); OXIDOREDUCTASE, SHORT-CHAIN DEHYDROGENASE/REDUCTASE SUPERFAMILY, FATTY; HET: JA3, MRD, NAP, GLU; 1.8A {STAPHYLOCOCCUS AUREUS SUBSP. AUREUS N315}	100	2.9e-39	31.2	251	282
☐ 3	5I7S_A	Enoyl-[acyl-carrier-protein] reductase [NADH] (E.C.1.3.1.9); enoyl-ACP reductase, Burkholderia pseudomallei, diphenyl; HET: E9P, NAD; 1.595A {Burkholderia pseudomallei} SCOP: c.2.1.2	100	5.3e-39	31.4	250	276
☐ 4	4ZJU_A	Enoyl-[acyl-carrier-protein] reductase [NADH] (E.C.1.3.1.9); SSGCID, NADH-dependent enoyl-ACP reductase, NAD; HET: NAD; 1.2A {Acinetobacter baumannii} SCOP: c.2.1.2	100	3.4e-39	30.1	256	275
☐ 5	4M89_A	Enoyl-[acyl-carrier-protein] reductase [NADH] (E.C.1.3.1.9); Enoyl, acyl carrier protein, ACP; HET: TCL, NAD; 1.9A {Neisseria meningitidis} SCOP: c.2.1.2	100	1.8e-38	31.7	255	261
☐ 6	2PD4_D	Enoyl-[acyl-carrier-protein] reductase [NADH] (E.C.1.3.1.9); antibacterial target, Helicobacter pylori, type; HET: NAD, DCN; 2.3A {Helicobacter pylori} SCOP: c.2.1.2	100	2.3e-38	31.5	257	275

Figure 14: Results of HHPred homologue matching

In Fig 14 it can be seen the results and the chosen protein match (*4D44*) which has the lower e-value combined with the closest number of match columns in the HMM-HMM alignment (S-S hits). The specific protein is the Crystal structure of S. aureus FabI in complex with NADP and 5-ethyl- 4-fluoro-2-((2-fluoropyridin-3-yl)oxy)phenol.

Question b

Now using both files doing α῾λυσταλΩ sequence analysis between *4D44*, *1c14* and the results are shown in Fig .15

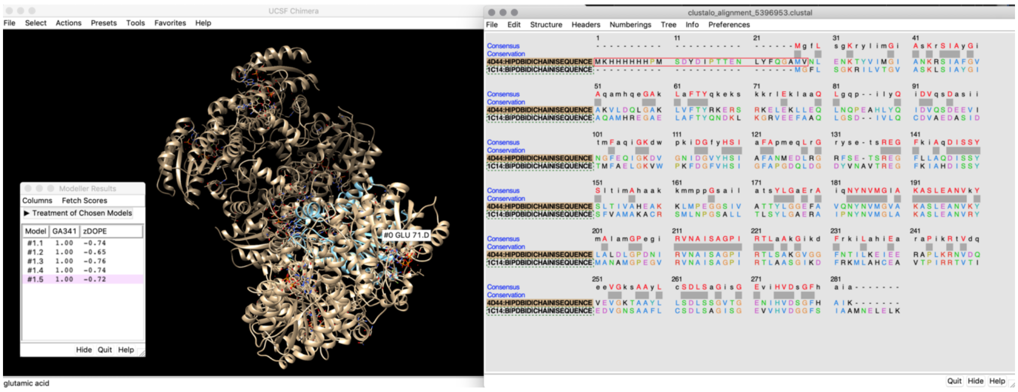


Number of sequences: **2**

```
☐ 4D44:H|PDBID|CHAIN|S   MKHHHHHHPMSDYDIPTTENLYFQGAMVNLENKTYVIMGIANKRSIAFGVAKVLDQLGAKLVFTYRKERSRKELEKLLEQLNQPE
☐ 1C14:B|PDBID|CHAIN|S   --------------------------MGFLSGKRILVTGVASKLSIAYGIAQAMHREGAELAFTYQNDKLKGRVEEFAAQLGSD-

☐ 4D44:H|PDBID|CHAIN|S   AHLYQIDVQSDEEVINGFEQIGKDVGNIDGVYHSIAFANMEDLRGRFSE-TSREGFLLAQDISSYSLTIVAHEAKKLMPEGGSIV
☐ 1C14:B|PDBID|CHAIN|S   -IVLQCDVAEDASIDTMFAELGKVWPKFDGFVHSIGFAPGDQLDGDYVNAVTREGFKIAHDISSYSFVAMAKACRSMLNPGSALL

☐ 4D44:H|PDBID|CHAIN|S   ATTYLGGEFAVQNYNVMGVAKASLEANVKYLALDLGPDNIRVNAISAGPIRTLSAKGVGGFNTILKEIEERAPLKRNVDQVEVGK
☐ 1C14:B|PDBID|CHAIN|S   TLSYLGAERAIPNYNVMGLAKASLEANVRYMANAMGPEGVRVNAISAGPIRTLAASGIKDFRKMLAHCEAVTPIRRTVTIEDVGN

☐ 4D44:H|PDBID|CHAIN|S   TAAYLLSDLSSGVTGENIHVDSGFHAIK-------
☐ 1C14:B|PDBID|CHAIN|S   SAAFLCSDLSAGISGEVVHVDGGFSIAAMNELELK
```

Figure 15: ClustalΩ sequence alignment

Using Chimera the PDB entry *1c14* was used and added H-bonds. The Clustal alignment was opened in Chimera and was used in order to build the homology models using the Modeller API.


UCSF Chimera
File Select Actions Presets Tools Favorites Help
#0 GLU 71.D
glutamic acid
Modeller Results
Columns Fetch Scores
▶ Treatment of Chosen Models
Hide Quit Help
clustalo_alignment_5396953.clustal
File Edit Structure Headers Numberings Tree Info Preferences
Quit Hide Help

Model	GA341	zDOPE
#1.1	1.00	-0.74
#1.2	1.00	-0.65
#1.3	1.00	-0.76
#1.4	1.00	-0.74
#1.5	1.00	-0.72

```
                           1           11          21          31          41
Consensus                  ----------  ----------  ------MgfL  sgKrylimGi  AsKrSIAyGi
Conservation
4D44:HIPDBIDICHAINISEQUENCE MKHHHHHHPM  SDYDIPTTEN  LYFQGAMVNL  ENKTYVIMGI  ANKRSIAFGV
1C14:BIPDBIDICHAINISEQUENCE ----------  ----------  ------MGFL  SGKRILVTGV  ASKLSIAYGI

                           51          61          71          81          91
Consensus                  AqamhqeGAk  LaFTYqkeks  kkrlEklaaQ  Lgqp--ilyQ  iDVqsDasii
Conservation
4D44:HIPDBIDICHAINISEQUENCE AKVLDQLGAK  LVFTYRKERS  RKELEKLLEQ  LNQPEAHLYQ  IDVQSDEEVI
1C14:BIPDBIDICHAINISEQUENCE AQAMHREGAE  LAFTYQNDKL  KGRVEEFAAQ  LGSD--IVLQ  CDVAEDASID

                           101         111         121         131         141
Consensus                  tmFaqiGKdw  pkiDGfyHSI  aFApmeqLrG  ryse-tsREG  FkiAqDISSY
Conservation
4D44:HIPDBIDICHAINISEQUENCE NGFEQIGKDV  GNIDGVYHSI  AFANMEDLRG  RFSE-TSREG  FLLAQDISSY
1C14:BIPDBIDICHAINISEQUENCE TMFAELGKVW  PKFDGFVHSI  GFAPGDQLDG  DYVNAVTREG  FKIAHDISSY

                           151         161         171         181         191
Consensus                  SltimAhaak  kmmppGsail  atsYLGaErA  iqNYNVMGIA  KASLEANVkY
Conservation
4D44:HIPDBIDICHAINISEQUENCE SLTIVAHEAK  KLMPEGGSIV  ATTYLGGEFA  VQNYNVMGVA  KASLEANVKY
1C14:BIPDBIDICHAINISEQUENCE SFVAMAKACR  SMLNPGSALL  TLSYLGAERA  IPNYNVMGLA  KASLEANVRY

                           201         211         221         231         241
Consensus                  mAlamGPegi  RVNAISAGPI  RTLaAkGikd  FrkiLahiEa  raPikRtVdq
Conservation
4D44:HIPDBIDICHAINISEQUENCE LALDLGPDNI  RVNAISAGPI  RTLSAKGVGG  FNTILKEIEE  RAPLKRNVDQ
1C14:BIPDBIDICHAINISEQUENCE MANAMGPEGV  RVNAISAGPI  RTLAASGIKD  FRKMLAHCEA  VTPIRRTVTI

                           251         261         271         281
Consensus                  eeVGksAAyL  cSDLSaGisG  EviHVDsGFh  aia-------
Conservation
4D44:HIPDBIDICHAINISEQUENCE VEVGKTAAYL  LSDLSSGVTG  ENIHVDSGFH  AIK-------
1C14:BIPDBIDICHAINISEQUENCE EDVGNSAAFL  CSDLSAGISG  EVVHVDGGFS  IAAMNELELK
```

Figure 16: Z-dope score

The choice of the #1.3 model was performed based on the lowest Z-dope score of -0.76 as shown in Fig 16 and the final structure (excluding chains B-H) is shown in Fig 17.

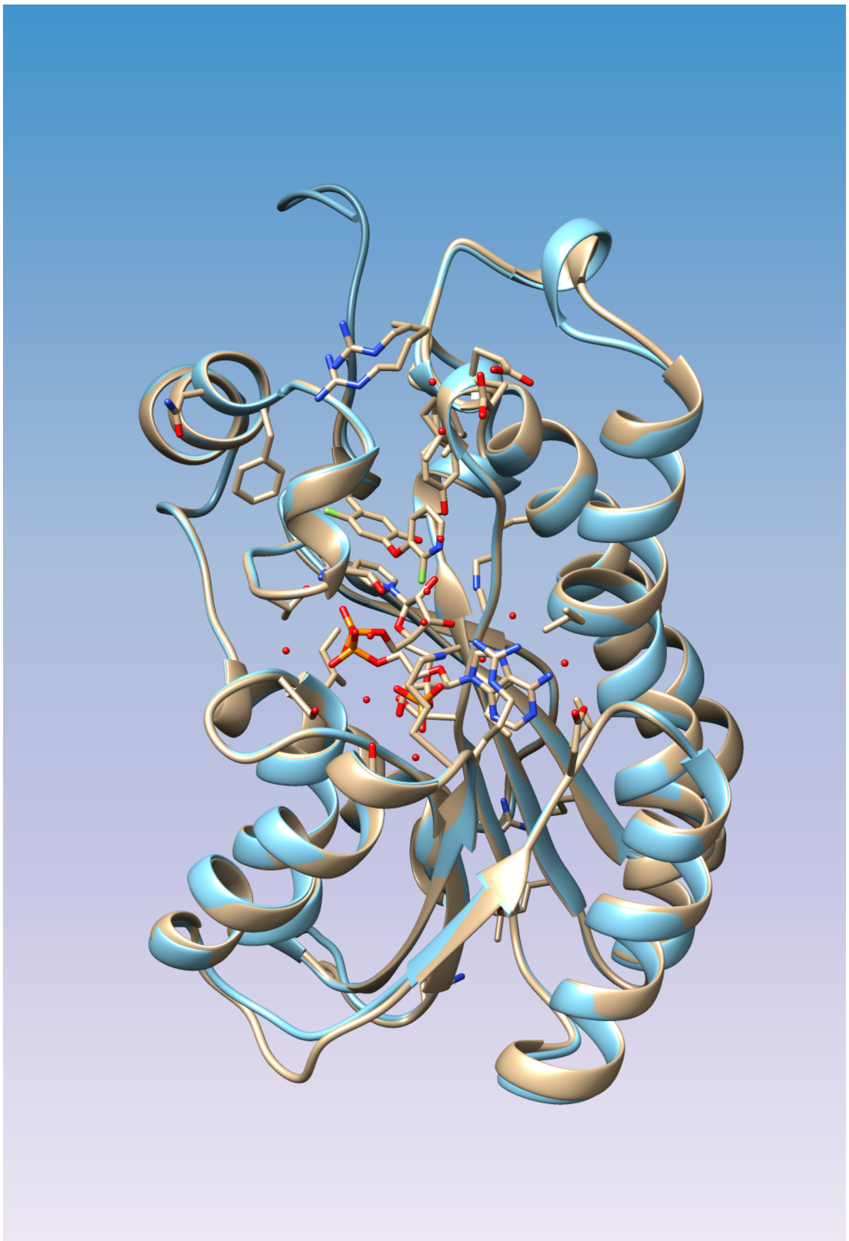
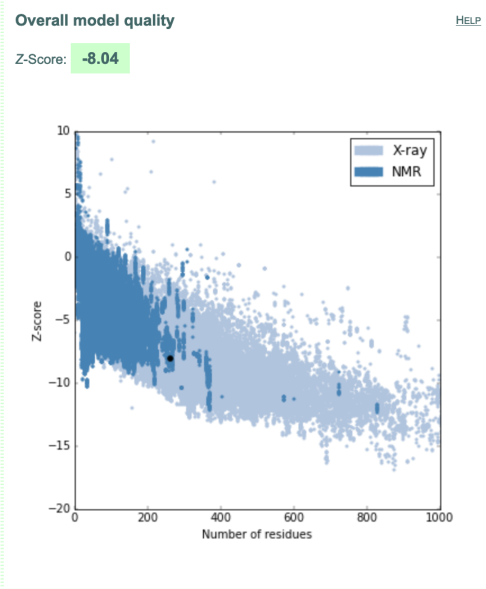


Figure 17: Final Chimera structure

To evaluate the stereochemical properties PROSA website is used and the results are shown in the following figures (Fig 18, Fig 19) Z-Score is -8.04 which seems to be on the range of scores typically found for the native structure (in fact the native was analysed above and was found to be -8.97, very similar indeed (see Fig 4).



Overall model quality

Z-Score: -8.04

Figure 18: Z-score for modelled protein

The local model quality looks also satisfactory as it is compact and thus scored like shown.

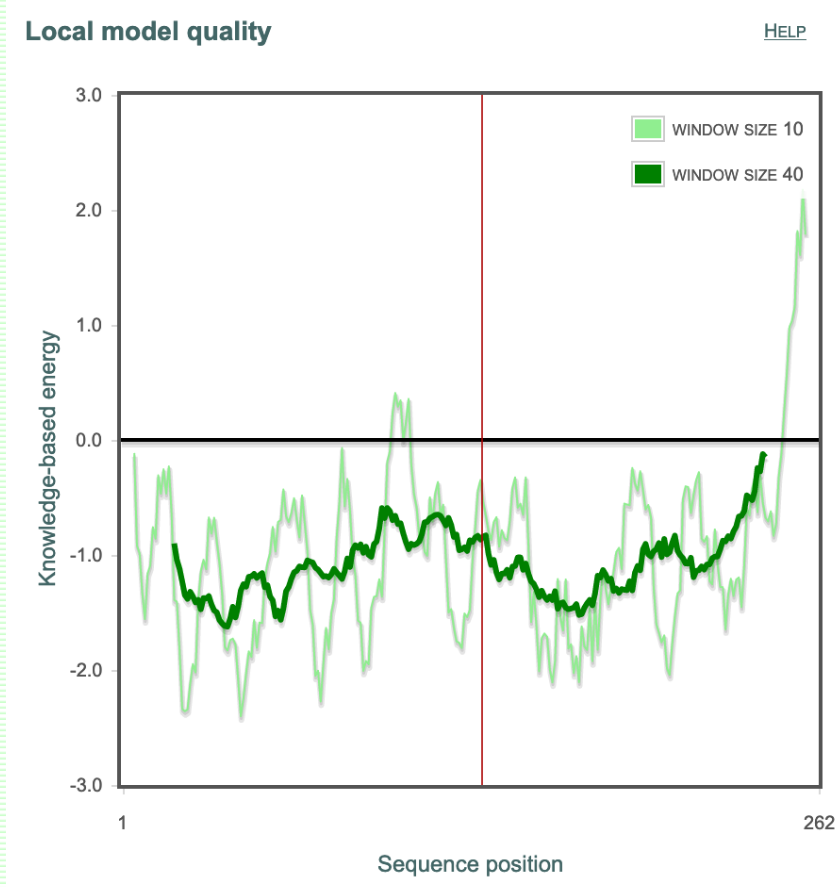


Figure 19: Local Model quality of modelled protein

Question c

For this part, using Chimera the Triclosan (TCL) is deleted from the 1c14 protein and the result is super-imposed to the modelled protein using a Chimera's MatchMaker function which is shown in Fig. 20.

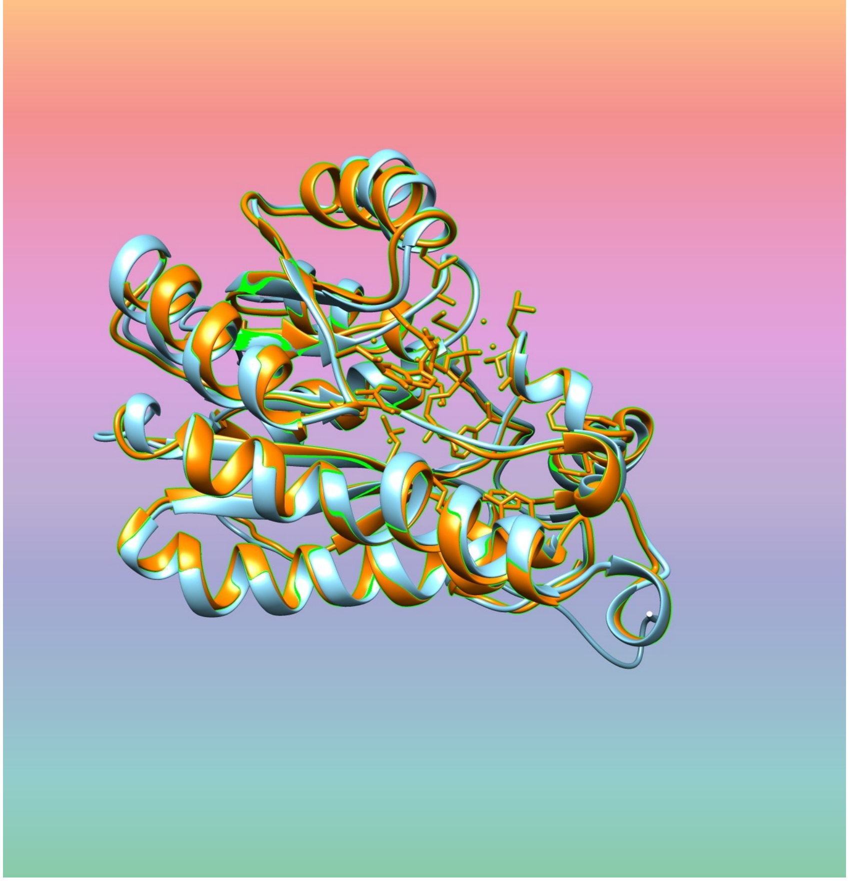
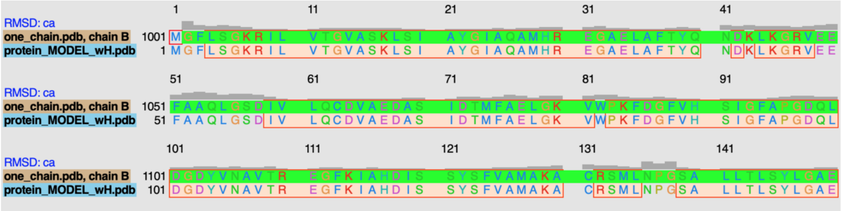


Figure 20: Super-imposed. Comparison of 1c14 minus TCL and modelled protein

A great similarity in the alignment of the sequences can be also seen in Fig. 21

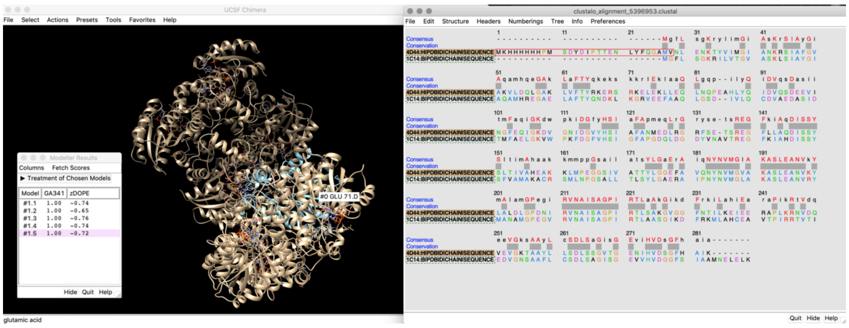


```
                                  1          11         21         31         41
RMSD: ca
one_chain.pdb, chain B    1001   MGFLSGKRIL VTGVASKLSI AYGIAQAMHR EGAELAFTYQ NDKLKGRVEE
protein_MODEL_wH.pdb         1   MGFLSGKRIL VTGVASKLSI AYGIAQAMHR EGAELAFTYQ NDKLKGRVEE

                                 51          61         71         81         91
RMSD: ca
one_chain.pdb, chain B    1051   FAAQLGSDIV LQCDVAEDAS IDTMFAELGK VWPKFDGFVH SIGFAPGDQL
protein_MODEL_wH.pdb        51   FAAQLGSDIV LQCDVAEDAS IDTMFAELGK VWPKFDGFVH SIGFAPGDQL

                                101         111        121        131        141
RMSD: ca
one_chain.pdb, chain B    1101   DGDYVNAVTR EGFKIAHDIS SYSFVAMAKA CRSMLNPGSA LLTLSYLGAE
protein_MODEL_wH.pdb       101   DGDYVNAVTR EGFKIAHDIS SYSFVAMAKA CRSMLNPGSA LLTLSYLGAE
```

Figure 21: Sequence match between the two proteins

Question d

In that step, a combination of Autodock (Sousa *et al.*, 2006)Vina was used on the Google Cloud online instance in combination with Autodock tools local installation on a Mac OS 10.14.


UCSF Chimera
File Select Actions Presets Tools Favorites Help
#0 GLU 71.D
clustalo_alignment_5396953.clustal
File Edit Structure Headers Numberings Tree Info Preferences
Quit Hide Help

Modeller Results

Columns Fetch Scores

▶ Treatment of Chosen Models

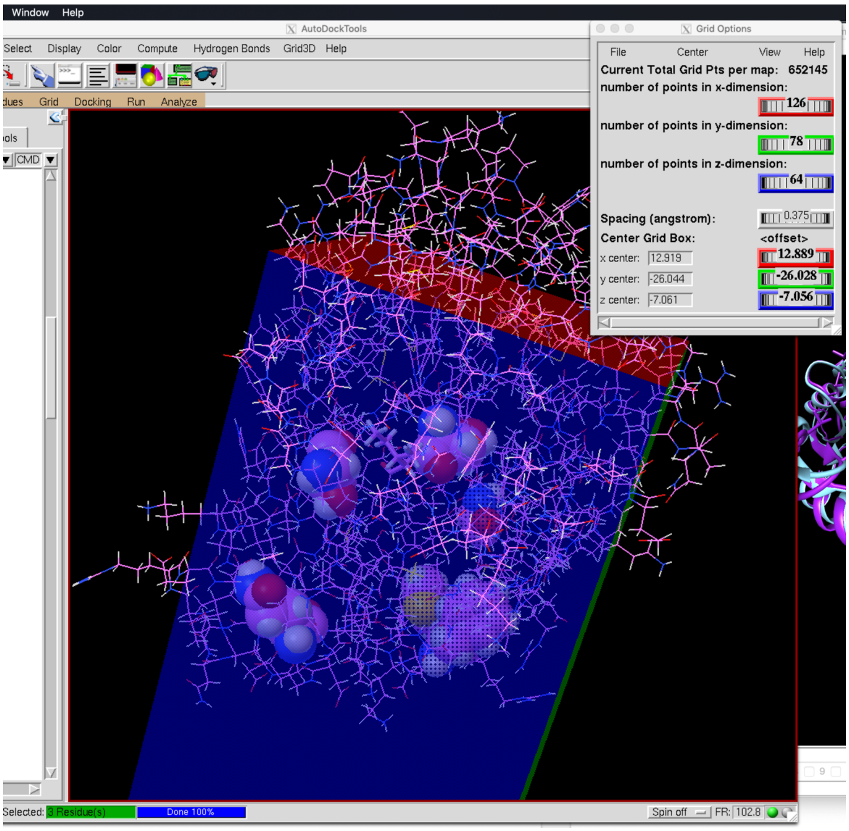
Model	GA341	zDOPE
#1.1	1.00	-0.74
#1.2	1.00	-0.65
#1.3	1.00	-0.76
#1.4	1.00	-0.74
#1.5	1.00	-0.72

Hide Quit Help

glutamic acid

Figure 22: Chimera ligand binding site prediction

The docking was performed by analysing in the most possible location of the ligand docking in Chimera, inputting those values in the Autodock Tools and using the settings to create a config file, for the command line docking, as shown in Fig 22 and Fig 23.


Window Help
AutoDockTools
Select Display Color Compute Hydrogen Bonds Grid3D Help
dues Grid Docking Run Analyze
ols
CMD
Spin off
FR: 102.8
9

Grid Options

File Center View Help

Current Total Grid Pts per map: 652145

number of points in x-dimension: 126

number of points in y-dimension: 78

number of points in z-dimension: 64

Spacing (angstrom): 0.375

Center Grid Box: <offset>

x center: 12.919 12.889

y center: -26.044 -26.028

z center: -7.061 -7.056

Selected: 3 Residue(s) Done 100%

Figure 23: Autodock tools set-up

Both of the ligand and protein files had to be converted in the *pdbqtp* file format using Autodock tools. The Autodock Vina was then executed from the command line using the following configuration file and commands,

```
code@instance-1:~/ioannis$ cat conf.txt
receptor = protein.pdbqt
ligand = ligand.pdbqt

center_x = 12.919
```

```
center_y = -26.044
center_z = -7.061

size_x = 126
size_y = 78
size_z = 64

exhaustiveness = 10
```

while the command line options given are shown here:

```
code@instance-1:~/ioannis$ ../bin/vina --config conf.txt --log outlog.log
```

That results is depicted in the output of the following table.

Using that table, the lowest affinity is -7.7 which presents the best fitting ligand to be docked in the requested protein model.

mode	affinity	dist from	best mode
	(kcal/mol)	rmsd l.b.	rmsd u.b.
1	-7.7	0.000	0.000
2	-7.6	1.312	5.879
3	-7.6	2.257	5.935
4	-7.0	2.280	2.988
5	-7.0	1.794	2.477
6	-6.7	14.552	16.153
7	-6.6	3.913	7.340
8	-6.2	10.032	11.737
9	-6.2	14.666	16.480

References

Current Protocols in Bioinformatics (2002) John Wiley & Sons Inc.

Bhargavi,M. *et al.* (2017) Identification of novel anti cancer agents by applying insilico methods for inhibition of TSPO protein.. *Comput Biol Chem*, **68**, 43–55.

Conte,L.L. (2000) SCOP: a Structural Classification of Proteins database. *Nucleic Acids Research*, **28**, 257–259.

Dasgupta,B. *et al.* (2007) Enhanced stability ofcisPro-Pro peptide bond in Pro-Pro-Phe sequence motif. *FEBS Letters*, **581**, 4529–4532.

Quirós,M. *et al.* (2018) Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database.. *J Cheminform*, **10**, 23.

Sousa,S.F. *et al.* (2006) Protein-ligand docking: Current status and future challenges. *Proteins: Structure Function, and Bioinformatics*, **65**, 15–26.

Wishart,D.S. *et al.* (2017) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, **46**, D1074–D1082.