DATE: 18-03-22

WEBLEM 3

Introduction to tertiary structure prediction

Proteins are involved in many cell activities (e.g., molecular transport, mechanical functions, message exchange) thus **knowing their 3D structure is crucial** in order to understand their function. **Protein tertiary structure prediction** is a research field which aims to **create models and software tools** able to predict the **three-dimensional shape of protein molecules** by describing the spatial disposition of each of its atoms starting from the sequence of its amino acids. There exist exact methods to **resolve the molecular structure with high precision**, but they are both time and resource consuming. **Computational based software techniques** can predict the tertiary structure of a protein with **acceptable precision** for many applications with high efficiency allowing for **genome-wide investigations**, otherwise not feasible.

Having a **computer-generated three-dimensional model** of a protein of interest has many ramifications, assuming it is reasonably correct. It may be of use for the **rational design of biochemical experiments**, such as **site-directed mutagenesis, protein stability, or functional analysis**. In addition to serving as a **theoretical guide to design experiments** for protein characterization, the model can help to **rationalize the experimental results** obtained with the protein of interest. In short, the modelling study helps to advance our **understanding of protein functions**.

METHODS:

There are **three computational approaches** to protein three-dimensional structural modelling and prediction. They are **homology modelling, threading,** and **ab initio prediction.**

HOMOLOGY MODELLING:

As the name suggests, **homology modelling** predicts protein structures based on **sequence homology** with known structures. It is also known as comparative modelling. The principle behind it is that if two proteins share a **high enough sequence similarity**, they are likely to have very **similar three-dimensional structures**. If one of the protein sequences has a **known structure**, then the structure **can be copied to the unknown protein** with a high degree of confidence. Homology modelling produces an **all-atom model** based on **alignment with template proteins.**

The overall homology modelling procedure consists of six steps.

1. **Template Selection** which involves identification of homologous sequences in the protein structure database to be used as templates for modelling
2. **Alignment** of target and template sequences.
3. **Building a framework structure** for the target protein consisting of main chain atoms.
4. **Refine and optimize** the entire model according to energy criteria.
5. **Evaluation** of the overall quality of the model obtained.

A number of **comprehensive modelling programs** are able to perform the complete procedure of homology modelling in an automated fashion. The **automation requires assembling a pipeline** that includes target **selection, alignment, model generation, and model evaluation.**

MODELLER:

MODELLER is a computer program for **comparative protein structure modelling**. In the simplest case, the input is an **alignment of a sequence** to be modelled with the **template structures**, the atomic coordinates of the templates, and a simple script file. **MODELLER** then automatically **calculates a model** containing all **non-hydrogen atoms**, within minutes on a modern PC and with no user intervention. Apart from model building, **MODELLER** can perform additional auxiliary tasks, including **fold assignment**, **alignment of two**

protein sequences or their profiles, **multiple alignment** of protein sequences and/or structures, **calculation of phylogenetic trees**, and **de novo modelling of loops** in protein structures.

THREADING AND FOLD RECOGNITION:

There are only **small number of protein folds available** (<1,000), compared to millions of protein sequences. This means that protein structures tend to be **more conserved** than protein sequences. Consequently, many proteins can share a **similar fold** even in the absence of **sequence similarities**. This allowed the development of computational methods to predict protein structures **beyond sequence similarities**. To determine whether a **protein sequence adopts** a known **three-dimensional structure** fold relies on **threading and fold recognition** methods. By definition, threading or structural fold recognition predicts the **structural fold** of an **unknown protein sequence** by fitting the sequence into a **structural database** and selecting the **best-fitting fold**. The comparison emphasizes matching of **secondary structures**, which are most evolutionarily conserved. Therefore, this approach can **identify structurally similar proteins** even without detectable sequence similarity.

The algorithms can be classified into two categories, **pairwise energy based** and **profile based**. The pairwise energy–based method was originally referred to as **threading** and the profile-based method was originally defined as **fold recognition**. However, the two terms are now often used **interchangeably without distinction** in the literature. A number of threading and fold recognition programs are available using **either or both prediction strategies.**

I-TASSER:

I-TASSER server is an on-line platform that implements the **I-TASSER based algorithms** for protein structure and function predictions. It allows academic users to **automatically generate high-quality model predictions** of 3D structure and **biological function** of protein molecules from their amino acid sequences**.** When user submits an amino acid sequence, the server **first tries to retrieve template proteins of similar folds** (or super-secondary structures) from the PDB library by LOMETS, a locally installed meta-threading approach.

In the **second step**, the continuous fragments excised from the PDB templates are reassembled into full-length models by **replica-exchange Monte Carlo simulations** with the threading unaligned regions (mainly loops) built by **ab initio modelling**. In cases where no appropriate template is identified by **LOMETS, I-TASSER** will build the whole structures by ab initio modelling. The low free-energy states are identified by **SPICKER** through clustering the simulation decoys.

In the **third step**, the fragment assembly simulation is performed again starting from the **SPICKER cluster centroids**, where the spatial restrains collected from **both** the **LOMETS** templates and the **PDB** structures by TM-align are used to **guide the simulations**. The purpose of the **second iteration** is to **remove the steric clash** as well as to **refine the global topology** of the cluster centroids. The decoys generated in the second simulations are then clustered and the lowest energy structures are selected. The **final full-atomic models** are obtained by **REMO** which builds the atomic details from the selected I-TASSER decoys through the optimization of the **hydrogen-bonding network.**

For **predicting the biological function** of the protein, the I-TASSER server matches the **predicted 3D models** to the **proteins in 3 independent libraries** which consist of proteins of known **enzyme classification (EC) number**, **gene ontology (GO) vocabulary**, and **ligand-binding sites.** The final results of function predictions are deduced from the consensus of **top structural matches** with the function scores calculated based on the confidence score of the I-TASSER structural models, the **structural similarity** between model and templates as **evaluated by TM-score**, and the sequence identity in the structurally aligned regions.

1. **What is C-score?**

   C-score is a **confidence score** for estimating the quality of predicted models by I-TASSER. It is calculated based on the **significance of threading template alignments** and the **convergence parameters** of the

structure assembly simulations. C-score is typically in the **range of [-5,2],** where a C-score of **higher value** signifies a model with a **high confidence** and vice-versa.

2. What is TM-score?

TM-score is a recently proposed scale for measuring the **structural similarity between two structures**. The purpose of proposing TM-score is to **solve** the problem of **RMSD** which is sensitive to the local error. Because RMSD is an **average distance** of all residue pairs in **two structures**, a **local** error (e.g., a misorientation of the tail) will arise a **big RMSD value** although the global topology is correct. In TM-score, however, the **small distance** is **weighted stronger** than the **big distance** which makes the score **insensitive to the local modelling error**. A **TM-score >0.5** indicates a model of correct topology and a TM-score<0.17 means a random similarity. These cut-off does not depend on the protein length.

3. What is difference and relationship between C-score and TM-score?

TM-score (or RMSD) is a known **standard** for measuring **structural similarity** between two structures which are usually used to measure the **accuracy of structure modelling** when the native structure is known, while **C-score** is a metric that I-TASSER developed to **estimate the confidence of the modelling**. In case where the native structure is not known, it becomes necessary to predict the quality of the modelling prediction, i.e., what is the distance between the predicted model and the native structures? To answer this question, we tried **predicting** the **TM-score** and **RMSD** of the predicted models relative the native structures based on the **C-score**.

In a benchmark test set of 500 non-homologous proteins, we found that C-score is highly correlated with TM-score and RMSD. Correlation coefficient of C-score of the first model with TM-score to the native structure is 0.91, while the coefficient of C-score with RMSD to the native structure is 0.75. These data lay the base for the reliable prediction of the TM-score and RMSD using C-score. In the output section, I-TASSER only reports the quality prediction (TM-score and RMSD) for the first model, because it was found that the correlation between C-score and TM-score is weak for lower rank models. However, the C-score is listed for all models just for a reference.

AB INITIO PROTEIN STRUCTURAL PREDICTION

The limited knowledge of protein folding forms the basis of ab initio prediction. As the name suggests, the ab initio prediction method attempts to produce all-atom protein models based on sequence information alone without the aid of known protein structures. The perceived advantage of this method is that predictions are not restricted by known folds and that novel protein folds can be identified. However, because the physicochemical laws governing protein folding are not yet well understood, the energy functions used in the ab initio prediction are at present rather inaccurate. The folding problem remains one of the greatest challenges in bioinformatics today.

Current ab initio algorithms are not yet able to accurately simulate the protein folding process. They work by using some type of heuristics. Because the native state of a protein structure is near energy minimum, the prediction programs are thus designed using the energy minimization principle. These algorithms search for every possible conformation to find the one with the lowest global energy. However, searching for a fold with the absolute minimum energy may not be valid in reality. This contributes to one of the fundamental flaws of this approach. In addition, searching for all possible structural conformations is not yet computationally feasible. It has been estimated that, by using one of the world's fastest supercomputers (one trillion operations per second), it takes 10 20 years to sample all possible conformations of a 40-residue protein. Therefore, some type of heuristics must be used to reduce the conformational space to be searched. Some recent ab initio methods combine fragment search and threading to yield a model of an unknown protein. The following web program is such an example using the hybrid approach.

ROBETTA:

The ROBETTA server provides automated tools for protein structure prediction and analysis. For structure prediction, sequences submitted to the server are parsed into putative domains and structural models are generated using either comparative modelling order novo structure prediction methods. If a confident match to a protein of known structure is found using BLAST, PSI-BLAST, FFAS03 or 3D-Jury, it is used as a template for comparative modelling. If no match is found, structure predictions are made using the de novo Rosetta fragment insertion method. Experimental nuclear magnetic resonance (NMR) constraints data can also be submitted with a query sequence for RosettaNMR de novo structure determination. Other current capabilities include the prediction of the effects of mutations on protein–protein interactions using computational interface alanine scanning. The Rosetta protein design and protein–protein docking methodologies will soon be available through the server as well.

INPUT AND OUTPUT:

Registration:

Users must register (http://robetta.bakerlab.org/register.jsp) before submitting jobs to Robetta.

Structure prediction server:

Sequences submitted to the structure prediction server must be in one-letter amino acid format. They can either be pasted into the submission form, or uploaded from a file. Users have the option to submit a sequence for either domain identification or full structure prediction. A user also has the option to specify the PDB id and chain for comparative modeling. For RosettaNMR submissions, a user must upload experimental NMR constraints data (chemical shifts, NOE data and/ or residual dipolar couplings). The required input format for each type of data is described at http://robetta.bakerlab.org/ documents/data_formats.jsp.

Results for a specific job are provided through the web interface by clicking on the job id listed in the queue table (http://robetta.bakerlab.org/queue.jsp). For full structure predictions, coordinates are also emailed to the user. For added insight, the following results are displayed along with the predicted models:

1. The prediction of transmembrane helices using TMHMM.
2. Low-complexity regions assigned by the program SEG
3. Coiled-coils prediction using COILS
4. The prediction of disordered regions using DISOPRED
5. Secondary structure predictions using PSIPRED, SAM-T99, Jufo and Jufo3D
6. The results listed above, domain predictions and the NR PSI-BLAST multiple sequence alignment used
7. For the last step in the domain prediction protocol condensed into an image to help corroborate the domain prediction results
8. Domain repeats prediction using REPRO predicted boundaries are given if repeats are detected
9. The top NR PSI-BLAST results and annotations for the top 20 species determined by lowest E-values.

The models for the full query are displayed as images at the bottom of the page. The coordinates for these models can be downloaded from the web site by clicking on the icons represented below each model image. Specific results are also provided for each domain by clicking on the domain number listed in the Ginzu domain prediction results table. For comparative models, the KSync alignment used for modelling is displayed. For de novo models, the Mammoth structure-model comparison results are displayed for the top 10 matches with Z-scores >4.5. The actual Mammoth structure-model alignment can be downloaded by clicking on the Z-score and viewed for further inspection using a molecular viewer such as RasMol. Users can download domain models by clicking on the icons below each domain model image.

Thus, modeller, I-TASSER and Robetta can be used to predict tertiary structures of proteins. These tools give faster results than x-ray or NMR techniques and can be used by researchers to understand protein functions.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 214-228.
2. Tradigo, Giuseppe (2018). Reference Module in Life Sciences || Algorithms for Structure Comparison and Analysis: Prediction of Tertiary Structures of Proteins. , (), –. doi:10.1016/B978-0-12-809633-8.20483-4
3. Bateman, Alex; Pearson, William R.; Stein, Lincoln D.; Stormo, Gary D.; Yates, John R. (2002). Current Protocols in Bioinformatics || Comparative Protein Structure Modeling Using MODELLER. , (), 5.6.1–5.6.37. doi:10.1002/cpbi.3
4. Tutorial. (n.d.). Salilab.org. Retrieved March 8, 2022, from https://salilab.org/modeller/tutorial/basic.html
5. I-TASSER server for protein structure and function prediction. (n.d.). Zhanggroup.org. Retrieved March 8, 2022, from https://zhanggroup.org/I-TASSER/about.html
6. Kim, D. E.; Chivian, D.; Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. , 32(0), 0–0. doi:10.1093/nar/gkh468

DATE: 18-03-22

WEBLEM 3a

MODELLER

(URL:https://salilab.org/modeller/)

AIM:

To perform tertiary structure prediction by comparative Modelling/Homology Modelling method using Modeller for query Rhodopsin

INTRODUCTION:

Rhodopsin, also called visual purple, pigment-containing sensory protein that converts light into an electrical signal. Rhodopsin is found in a wide range of organisms, from vertebrates to bacteria. In many seeing animals, including humans, it is required for vision in dim light and is located in the retina of the eye—specifically, within the tightly packed disks that make up the outer segment of the retina's photoreceptive rod cells, which are specially adapted for vision under low-light conditions.

MODELLER is a computer program for **comparative protein structure modelling**. In the simplest case, the input is an **alignment of a sequence** to be modelled with the **template structures**, the atomic coordinates of the templates, and a simple script file. **MODELLER** then automatically **calculates a model** containing all **non-hydrogen atoms**, within minutes on a modern PC and with no user intervention. Apart from model building, **MODELLER** can perform additional auxiliary tasks, including **fold assignment**, **alignment of two protein sequences** or their profiles, **multiple alignment** of protein sequences and/or structures, **calculation of phylogenetic trees**, and **de novo modelling of loops** in protein structures.

METHODOLOGY:

1. Install modeller. (URL: https://salilab.prg/modeller)
2. Retrieve FASTA sequence for enzyme rhodopsin
3. Follow the steps given in the tutorial section.
4. Run scripts for searching for structures related to query, selecting template target-template alignment and model building/
5. Observe and interpret the results.

OBSERVATION:

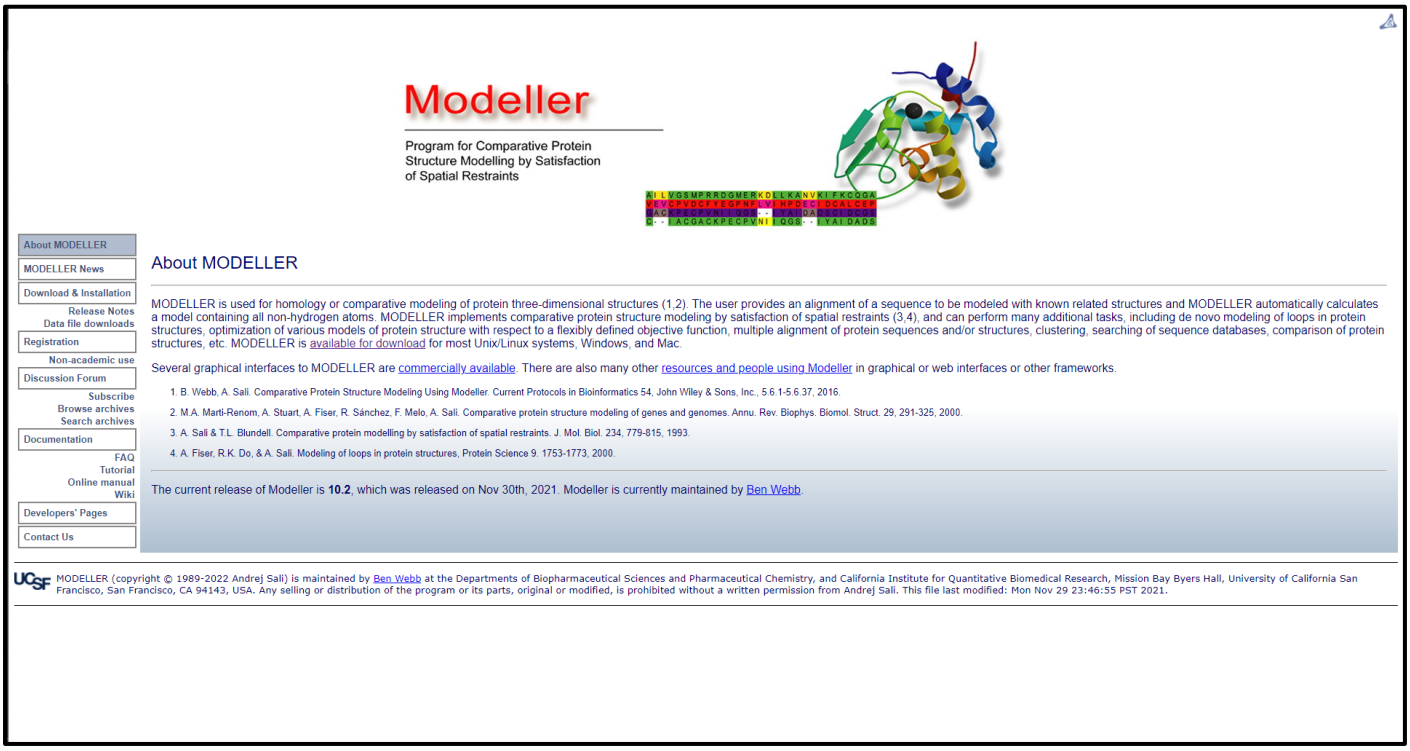


Fig1. Homepage for Modeller

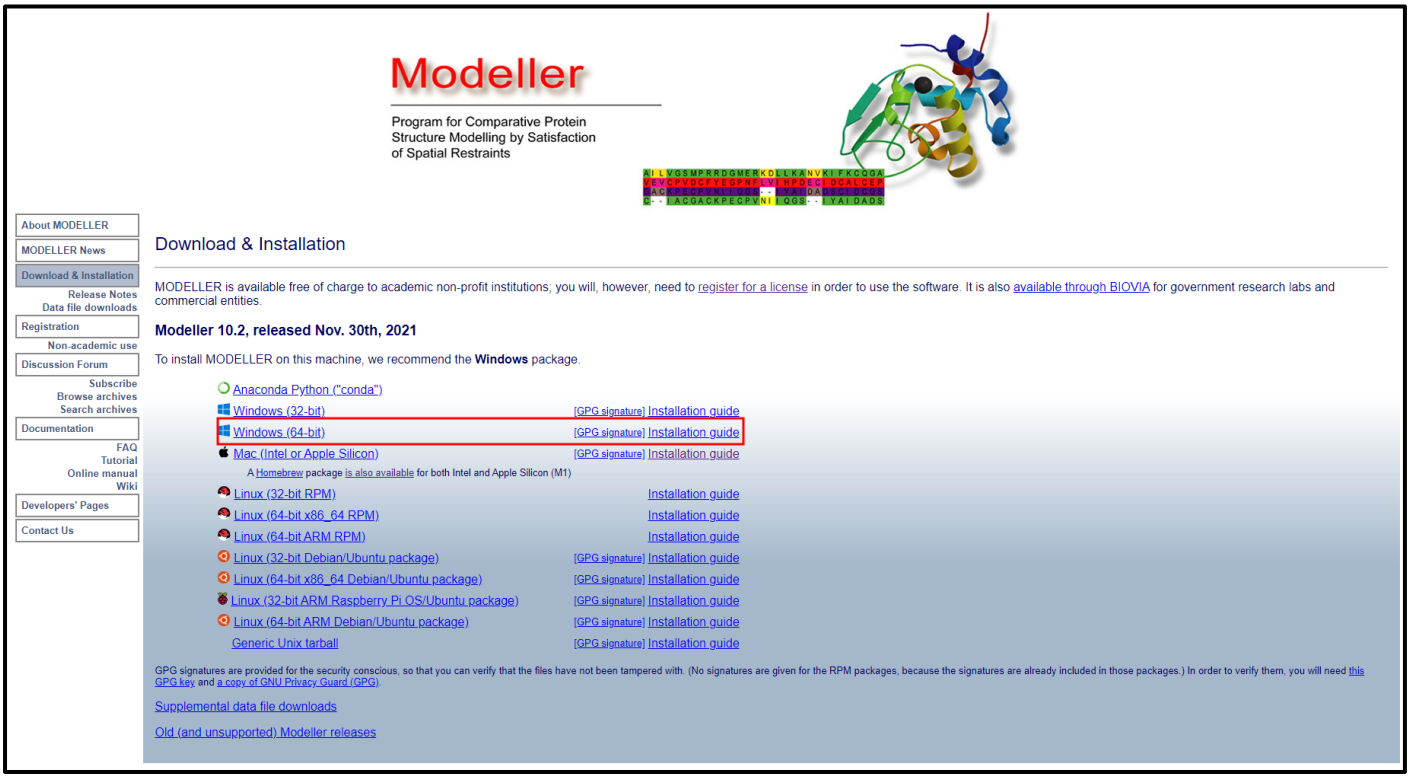


Fig2. Page to install Modeller

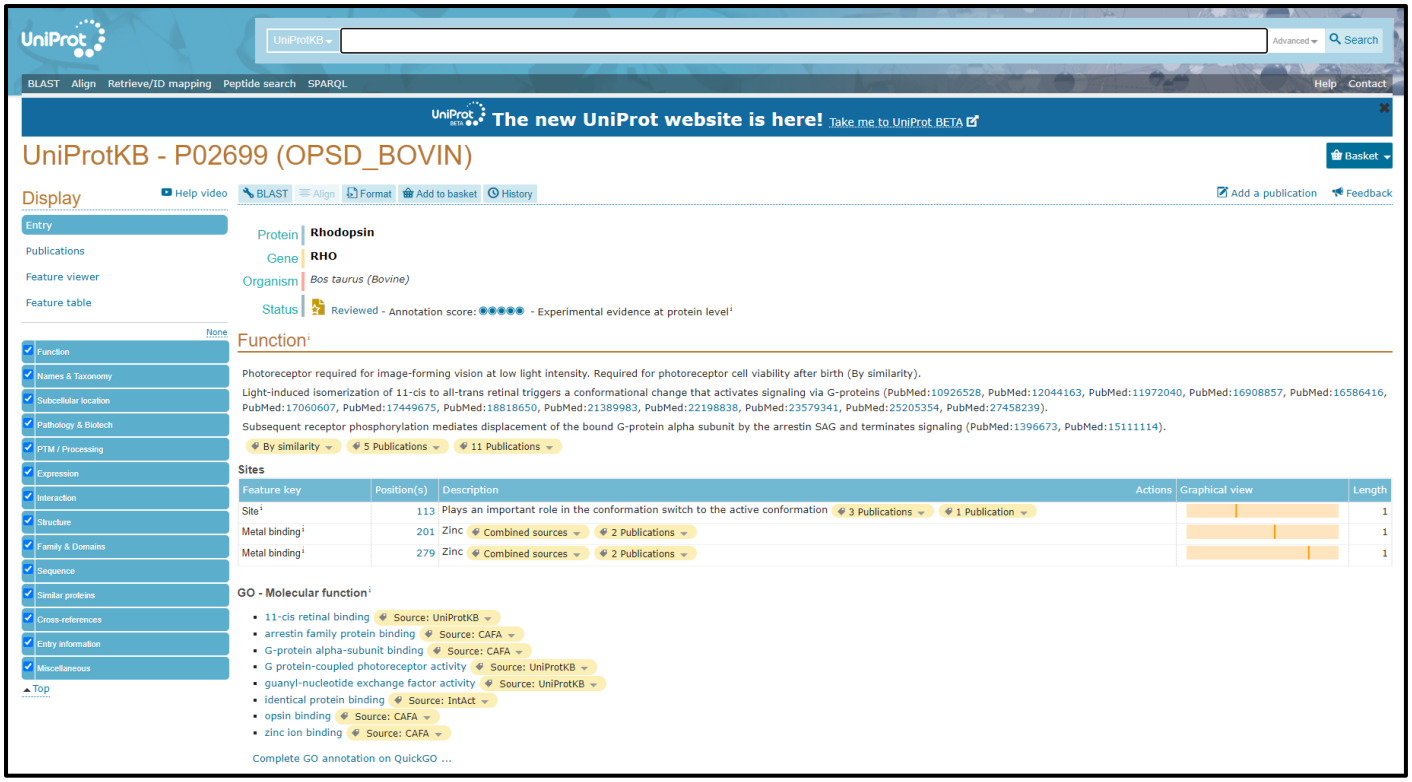


Fig3. Result page for Rhodopsin in UniProt database

```
>sp|Q15835|GRK1_HUMAN Rhodopsin kinase GRK1 OS=Homo sapiens OX=9606 GN=GRK1 PE=1
SV=1
MDFGSLETVVANSAFIAARGSFDGSSSQPSRDKKYLAKLKLPPLSKCESLRDSLSLEFES
VCLEQPIGKKLFQQFLQSAEKHLPALELWKDIEDYDTADNDLQPQKAQTILAQYLDPQAK
LFCSFLDEGIVAKFKEGPVEIQDGLFQPLLQATLAHLGQAPFQEYLGSLYFLRFLQWKWL
EAQPMGEDWFLDFRVLGKGGFGEVSACQMKATGKLYACKKLNKKRLKKRKGYQGAMVEKK
ILMKVHSRFIVSLAYAFETKADLCLVMTIMNGGDIRYHIYNVNEENPGFPEPRALFYTAQ
IICGLEHLHQRRIVYRDLKPENVLLDNDGNVRISDLGLAVELLDGQSKTKGYAGTPGFMA
PELLQGEEYDFSVDYFALGVTLYEMIAARGPFRARGEKVENKELKHRIISEPVKYPDKFS
QASKDFCEALLEKDPEKRLGFRDETCDKLRAHPLFKDLNWRQLEAGMLMPPFIPDSKTVY
AKDIQDVGAFSTVKGVAFDKTDTEFFQEFATGNCPIPWQEEMIETGIFGELNVWRSDGQM
PDDMKGISGGSSSSSKSGMCLVS
```

Fig4. FASTA sequence for Rhodopsin

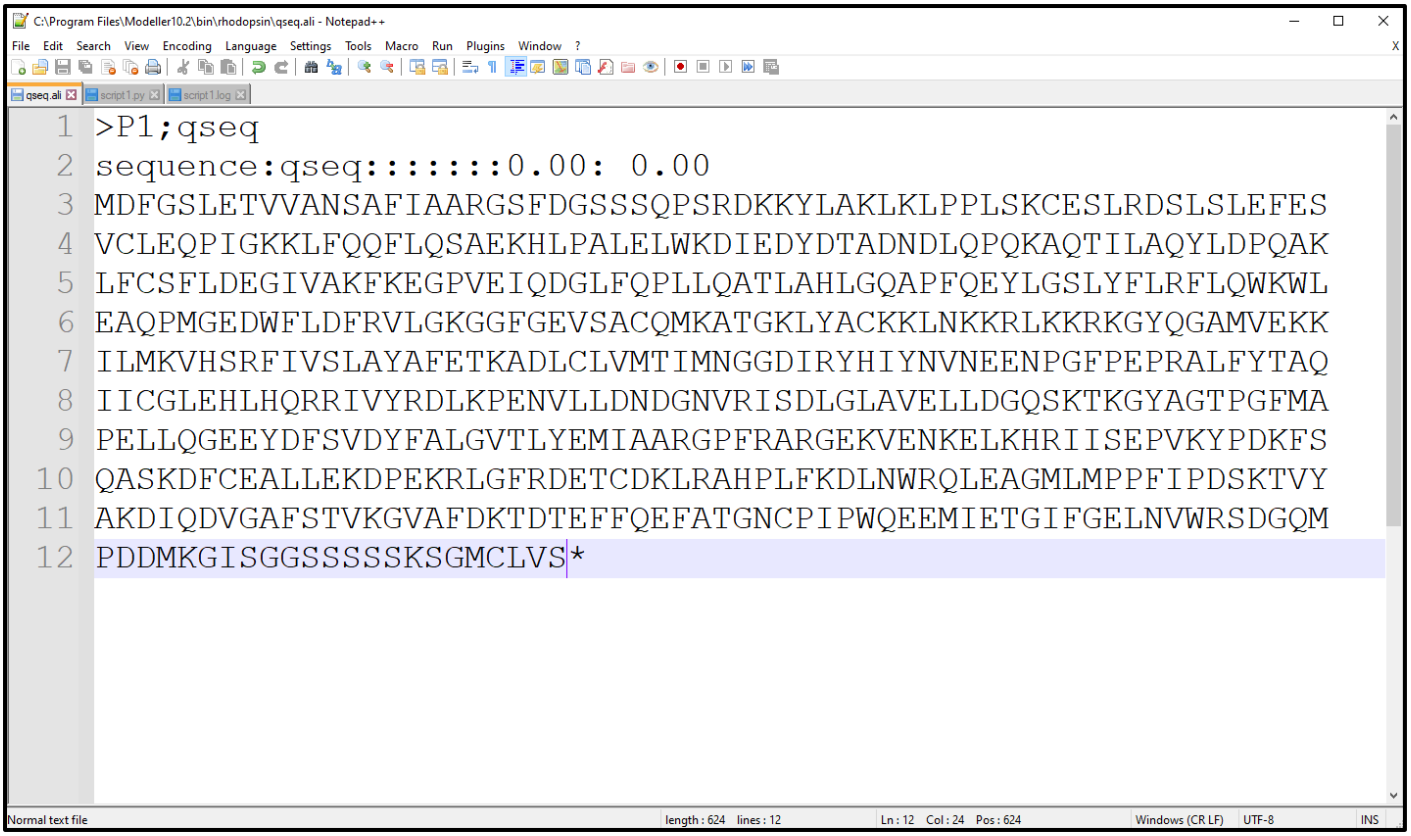


Fig5. FASTA sequence in PIR format

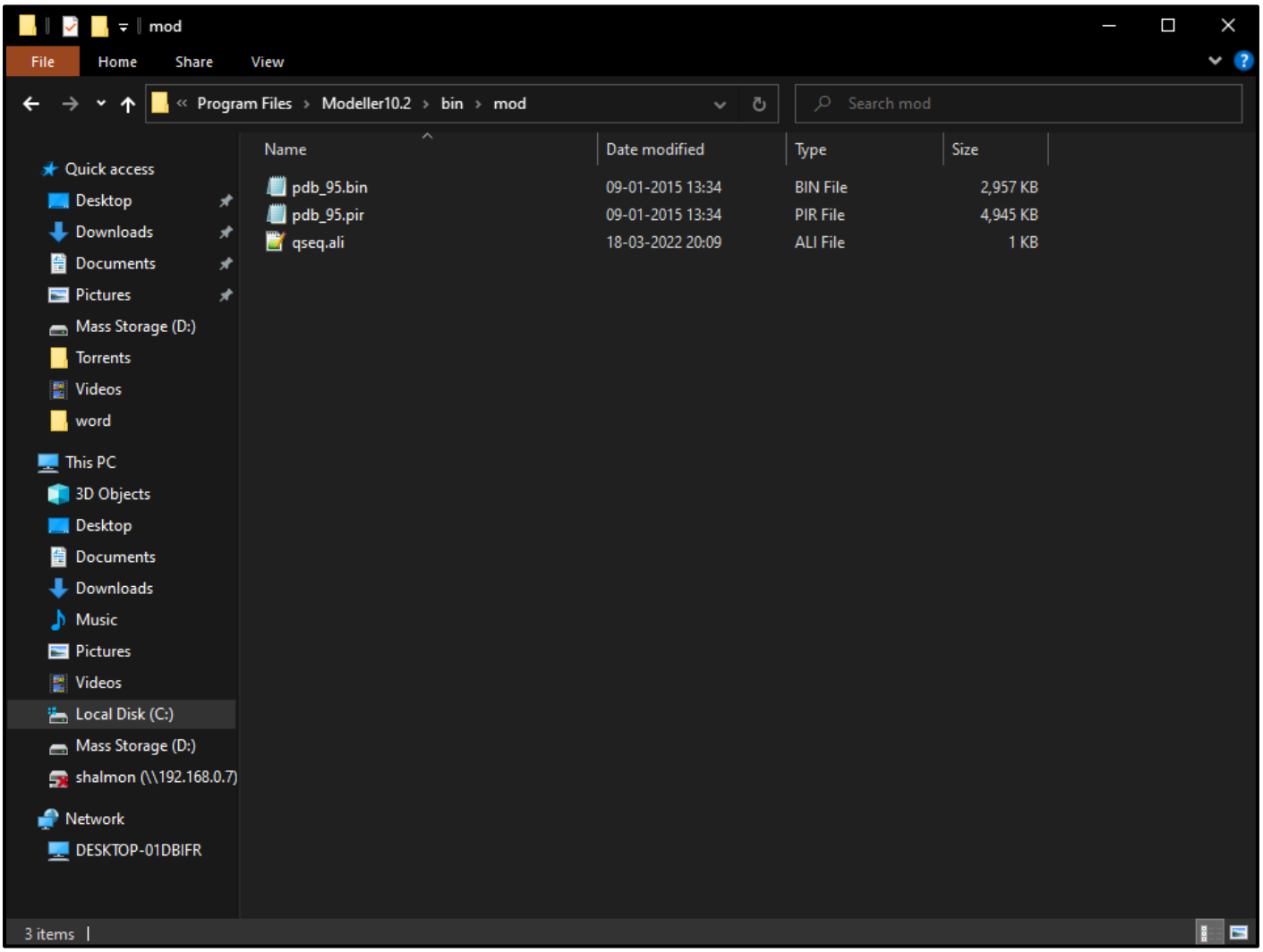


Fig6. Target sequence saved in .ali format

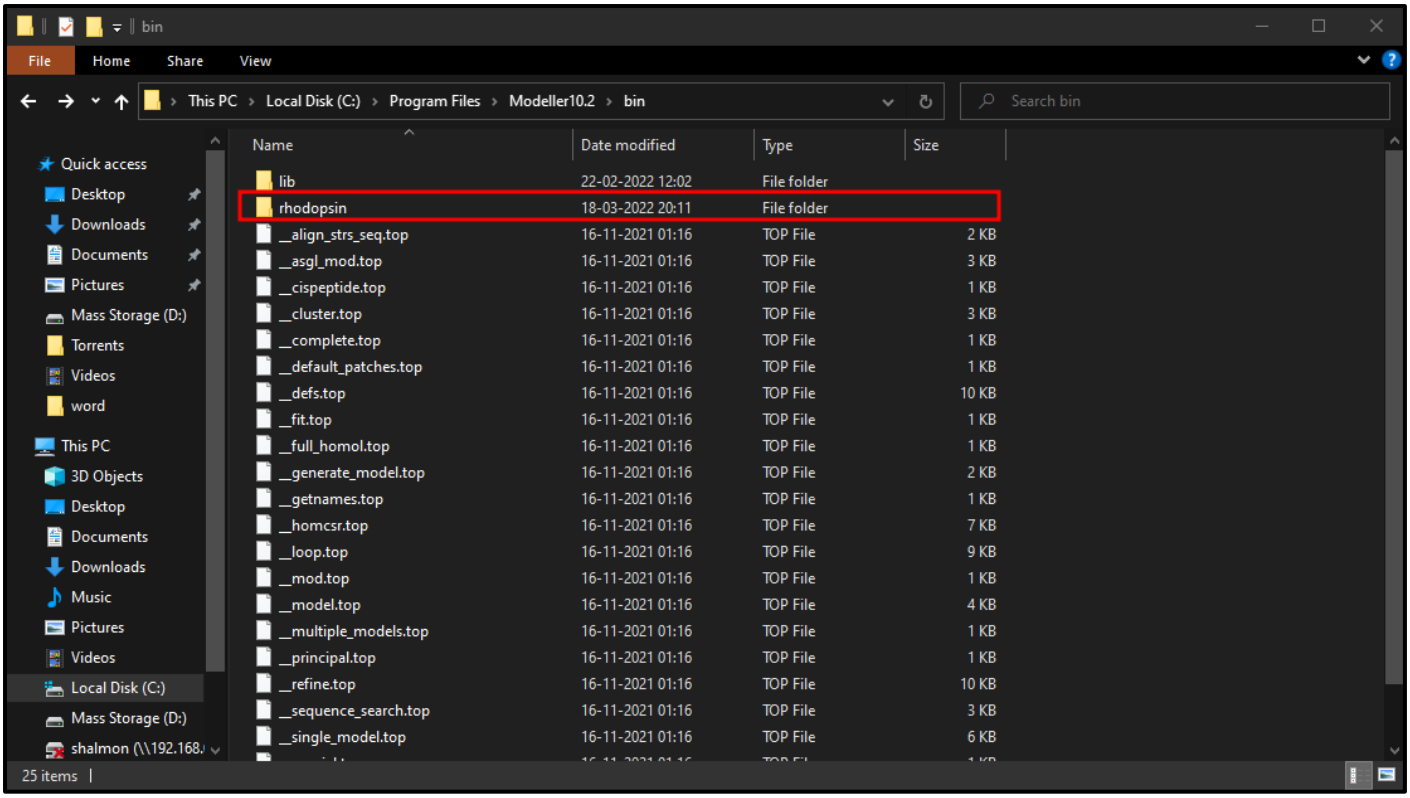


Fig7. Rhodopsin folder saved in the bin folder of modeller

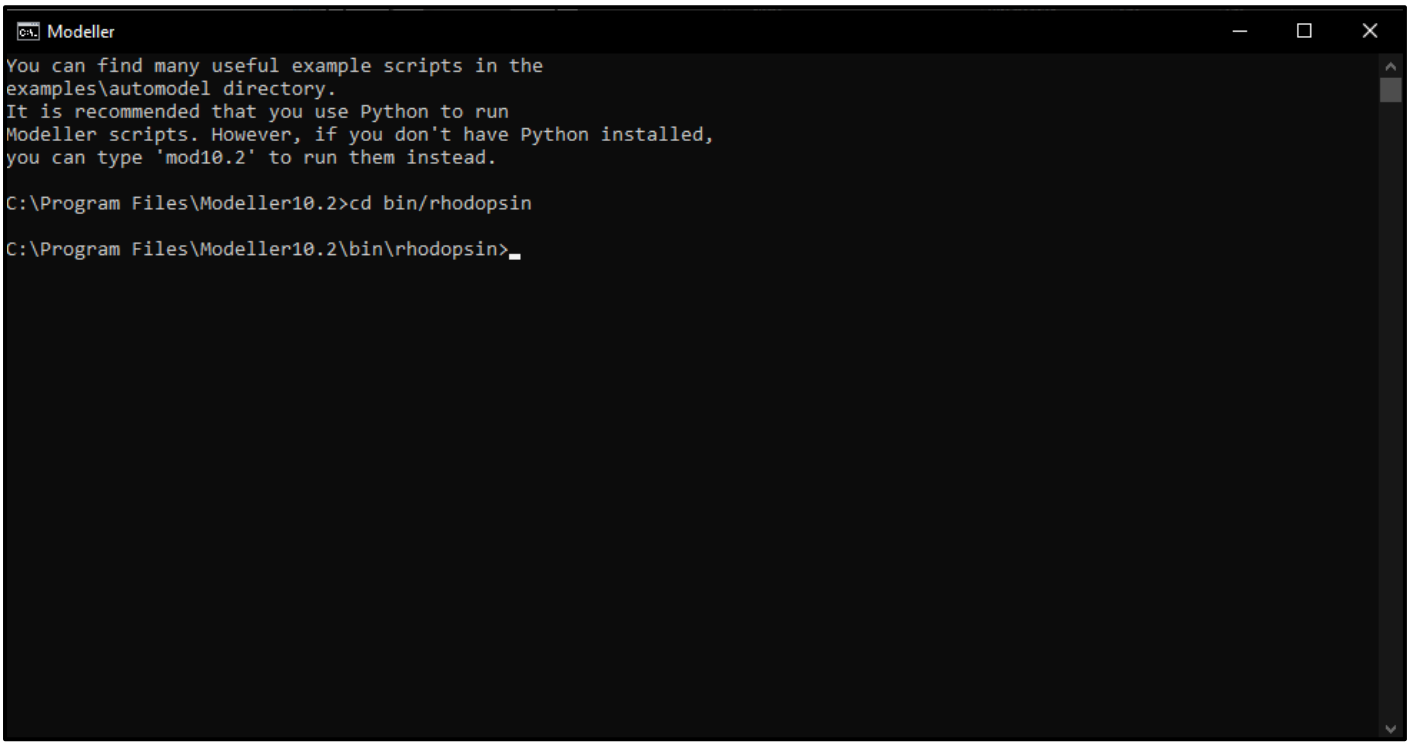


Fig8. Setting Working directory in Modeller command line

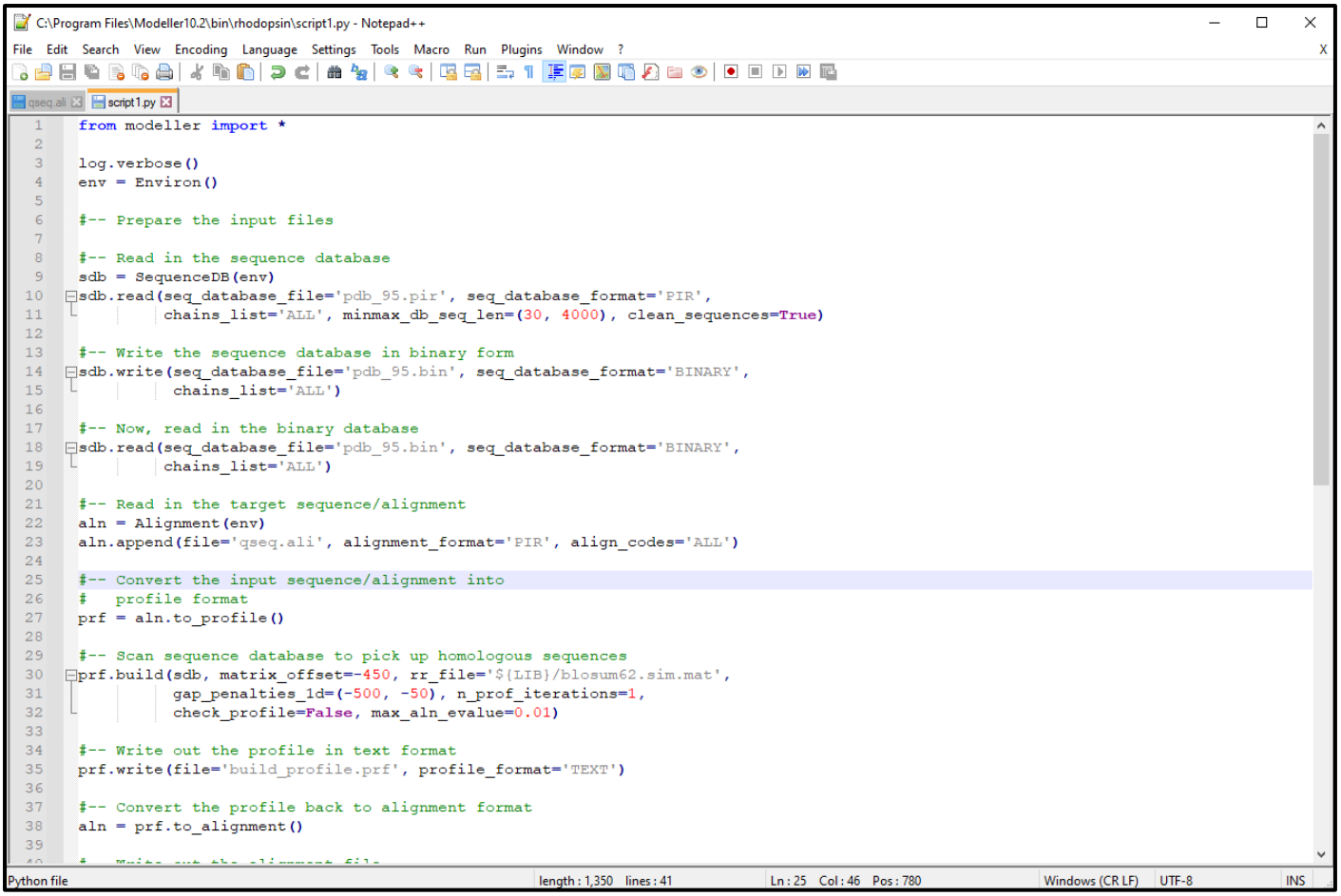


Fig9. Python script for searching for structures relation to rhodopsin

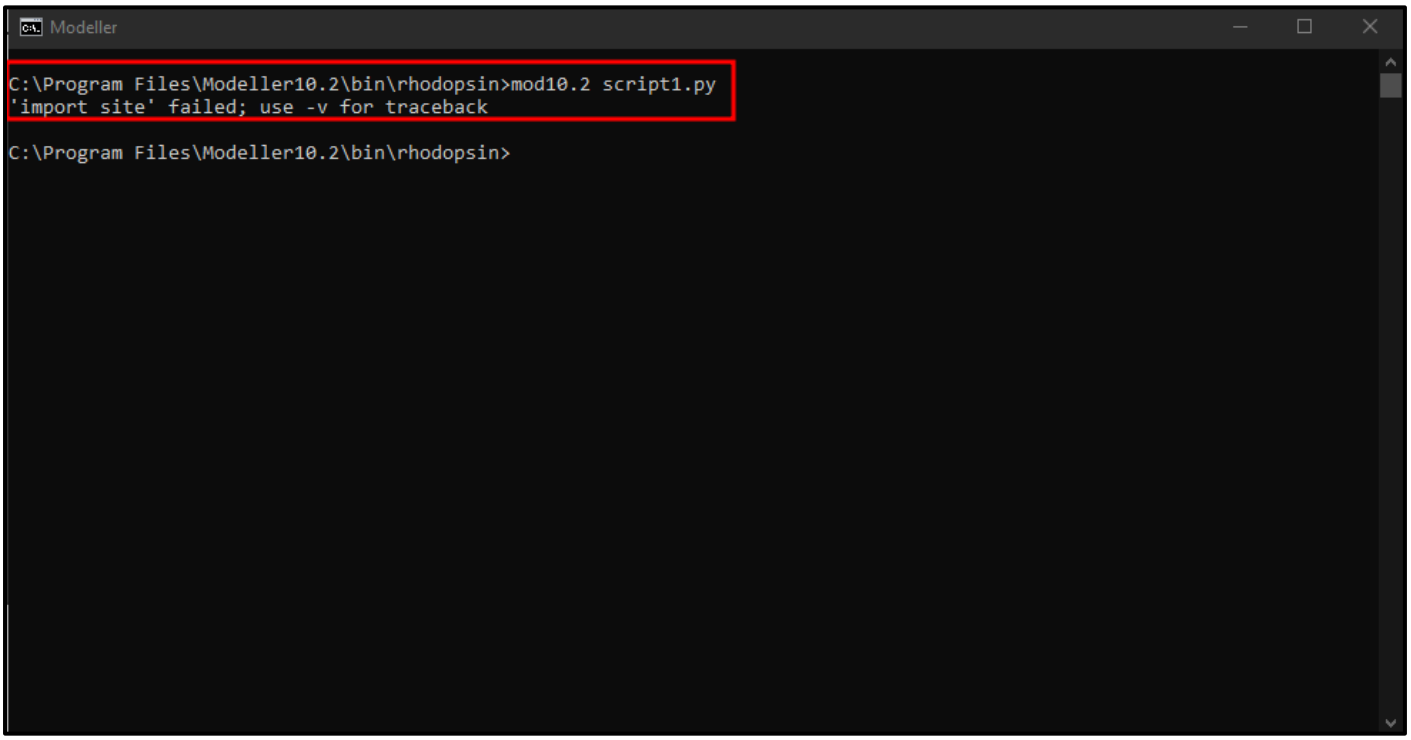
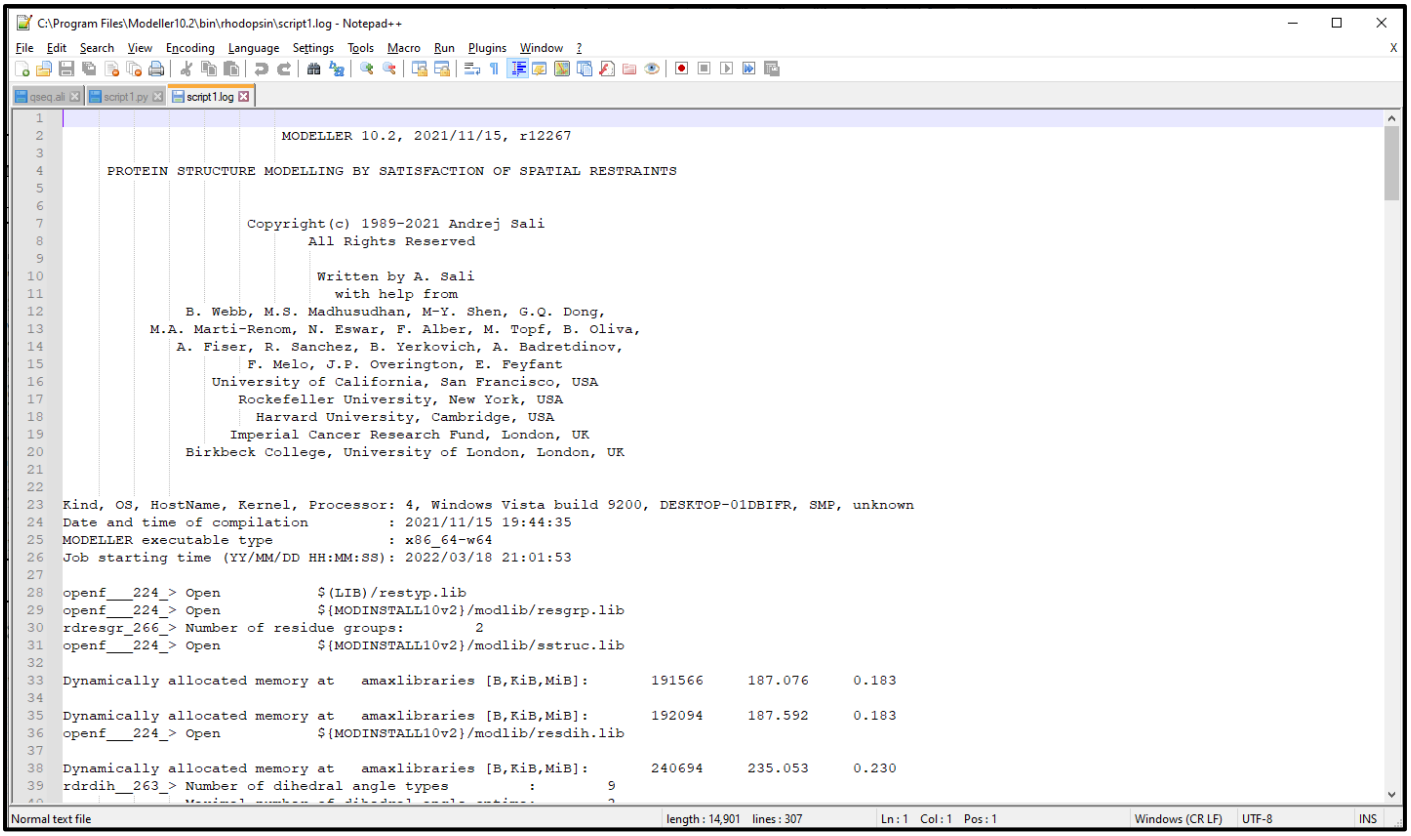
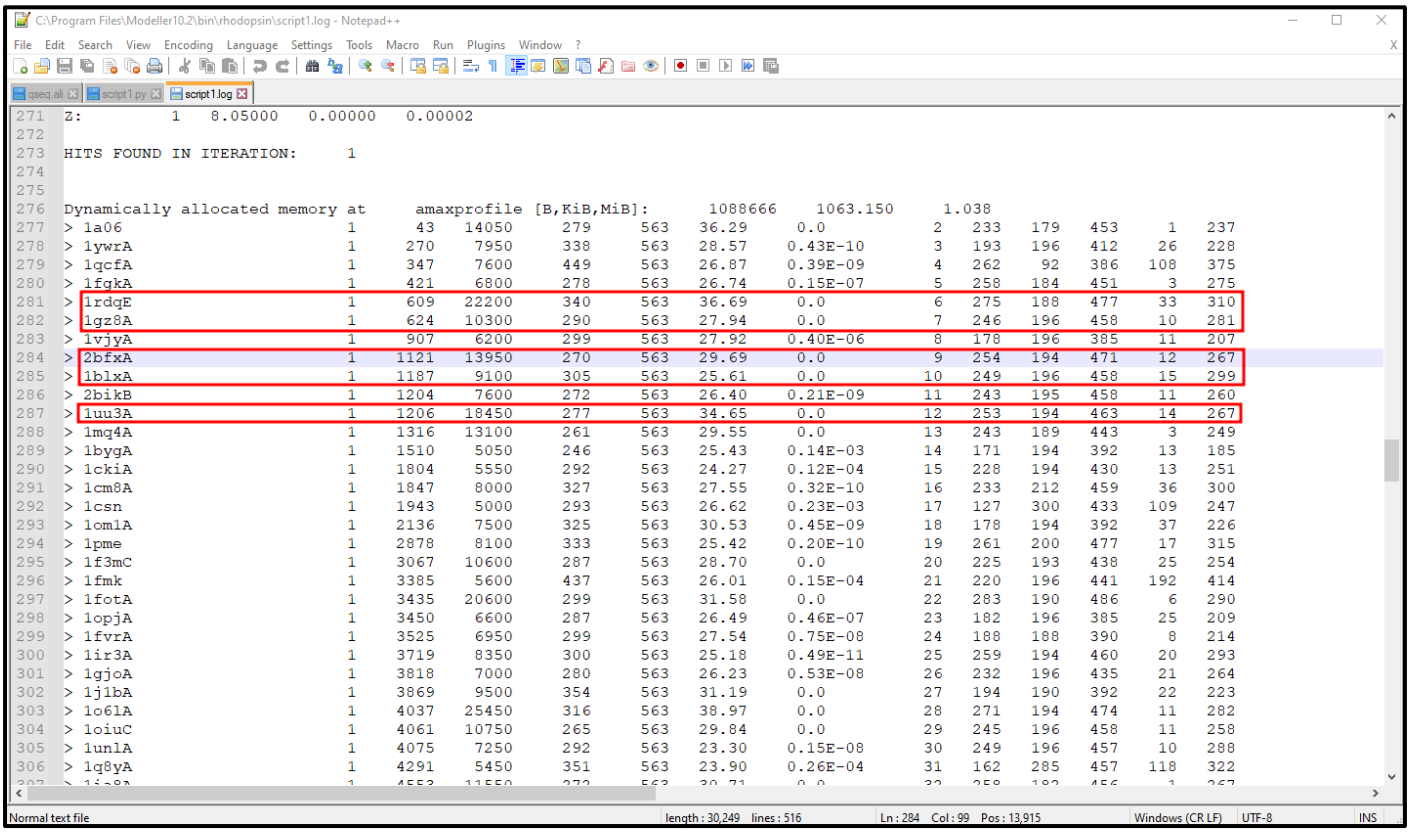


Fig10. Running script1.py



```
C:\Program Files\Modeller10.2\bin\rhodopsin\script1.log - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
qseq.ali | script1.py | script1.log

                    MODELLER 10.2, 2021/11/15, r12267

     PROTEIN STRUCTURE MODELLING BY SATISFACTION OF SPATIAL RESTRAINTS


                    Copyright(c) 1989-2021 Andrej Sali
                           All Rights Reserved

                            Written by A. Sali
                              with help from
                 B. Webb, M.S. Madhusudhan, M-Y. Shen, G.Q. Dong,
               M.A. Marti-Renom, N. Eswar, F. Alber, M. Topf, B. Oliva,
                  A. Fiser, R. Sanchez, B. Yerkovich, A. Badretdinov,
                          F. Melo, J.P. Overington, E. Feyfant
                     University of California, San Francisco, USA
                        Rockefeller University, New York, USA
                          Harvard University, Cambridge, USA
                       Imperial Cancer Research Fund, London, UK
                   Birkbeck College, University of London, London, UK


Kind, OS, HostName, Kernel, Processor: 4, Windows Vista build 9200, DESKTOP-01DBIFR, SMP, unknown
Date and time of compilation         : 2021/11/15 19:44:35
MODELLER executable type             : x86_64-w64
Job starting time (YY/MM/DD HH:MM:SS): 2022/03/18 21:01:53

openf___224_> Open           $(LIB)/restyp.lib
openf___224_> Open           ${MODINSTALL10v2}/modlib/resgrp.lib
rdresgr_266_> Number of residue groups:        2
openf___224_> Open           ${MODINSTALL10v2}/modlib/sstruc.lib

Dynamically allocated memory at   amaxlibraries [B,KiB,MiB]:       191566     187.076     0.183

Dynamically allocated memory at   amaxlibraries [B,KiB,MiB]:       192094     187.592     0.183
openf___224_> Open           ${MODINSTALL10v2}/modlib/resdih.lib

Dynamically allocated memory at   amaxlibraries [B,KiB,MiB]:       240694     235.053     0.230
rdrdih__263_> Number of dihedral angle types         :         9

Normal text file   length : 14,901   lines : 307   Ln : 1   Col : 1   Pos : 1   Windows (CR LF)   UTF-8   INS
```

Fig11. Log file for script1.py



```
C:\Program Files\Modeller10.2\bin\rhodopsin\script1.log - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
qseq.ali | script1.py | script1.log

Z:         1    8.05000    0.00000    0.00002

HITS FOUND IN ITERATION:      1


Dynamically allocated memory at     amaxprofile [B,KiB,MiB]:      1088666     1063.150     1.038
> 1a06                      1      43   14050      279      563   36.29     0.0           2   233   179   453     1   237
> 1ywrA                     1     270    7950      338      563   28.57     0.43E-10      3   193   196   412    26   228
> 1qcfA                     1     347    7600      449      563   26.87     0.39E-09      4   262    92   386   108   375
> 1fgkA                     1     421    6800      278      563   26.74     0.15E-07      5   258   184   451     3   275
> 1rdqE                     1     609   22200      340      563   36.69     0.0           6   275   188   477    33   310
> 1gz8A                     1     624   10300      290      563   27.94     0.0           7   246   196   458    10   281
> 1vjyA                     1     907    6200      299      563   27.92     0.40E-06      8   178   196   385    11   207
> 2bfxA                     1    1121   13950      270      563   29.69     0.0           9   254   194   471    12   267
> 1blxA                     1    1187    9100      305      563   25.61     0.0          10   249   196   458    15   299
> 2bikB                     1    1204    7600      272      563   26.40     0.21E-09     11   243   195   458    11   260
> 1uu3A                     1    1206   18450      277      563   34.65     0.0          12   253   194   463    14   267
> 1mq4A                     1    1316   13100      261      563   29.55     0.0          13   243   189   443     3   249
> 1bygA                     1    1510    5050      246      563   25.43     0.14E-03     14   171   194   392    13   185
> 1ckiA                     1    1804    5550      292      563   24.27     0.12E-04     15   228   194   430    13   251
> 1cm8A                     1    1847    8000      327      563   27.55     0.32E-10     16   233   212   459    36   300
> 1csn                      1    1943    5000      293      563   26.62     0.23E-03     17   127   300   433   109   247
> 1om1A                     1    2136    7500      325      563   30.53     0.45E-09     18   178   194   392    37   226
> 1pme                      1    2878    8100      333      563   25.42     0.20E-10     19   261   200   477    17   315
> 1f3mC                     1    3067   10600      287      563   28.70     0.0          20   225   193   438    25   254
> 1fmk                      1    3385    5600      437      563   26.01     0.15E-04     21   220   196   441   192   414
> 1fotA                     1    3435   20600      299      563   31.58     0.0          22   283   190   486     6   290
> 1opjA                     1    3450    6600      287      563   26.49     0.46E-07     23   182   196   385    25   209
> 1fvrA                     1    3525    6950      299      563   27.54     0.75E-08     24   188   188   390     8   214
> 1ir3A                     1    3719    8350      300      563   25.18     0.49E-11     25   259   194   460    20   293
> 1gjoA                     1    3818    7000      280      563   26.23     0.53E-08     26   232   196   435    21   264
> 1j1bA                     1    3869    9500      354      563   31.19     0.0          27   194   190   392    22   223
> 1o61A                     1    4037   25450      316      563   38.97     0.0          28   271   194   474    11   282
> 1oiuC                     1    4061   10750      265      563   29.84     0.0          29   245   196   458    11   258
> 1unlA                     1    4075    7250      292      563   23.30     0.15E-08     30   249   196   457    10   288
> 1q8yA                     1    4291    5450      351      563   23.90     0.26E-04     31   162   285   457   118   322

Normal text file   length : 30,249   lines : 516   Ln : 284   Col : 99   Pos : 13,915   Windows (CR LF)   UTF-8   INS
```

Fig11.1. Hits found for similar structures

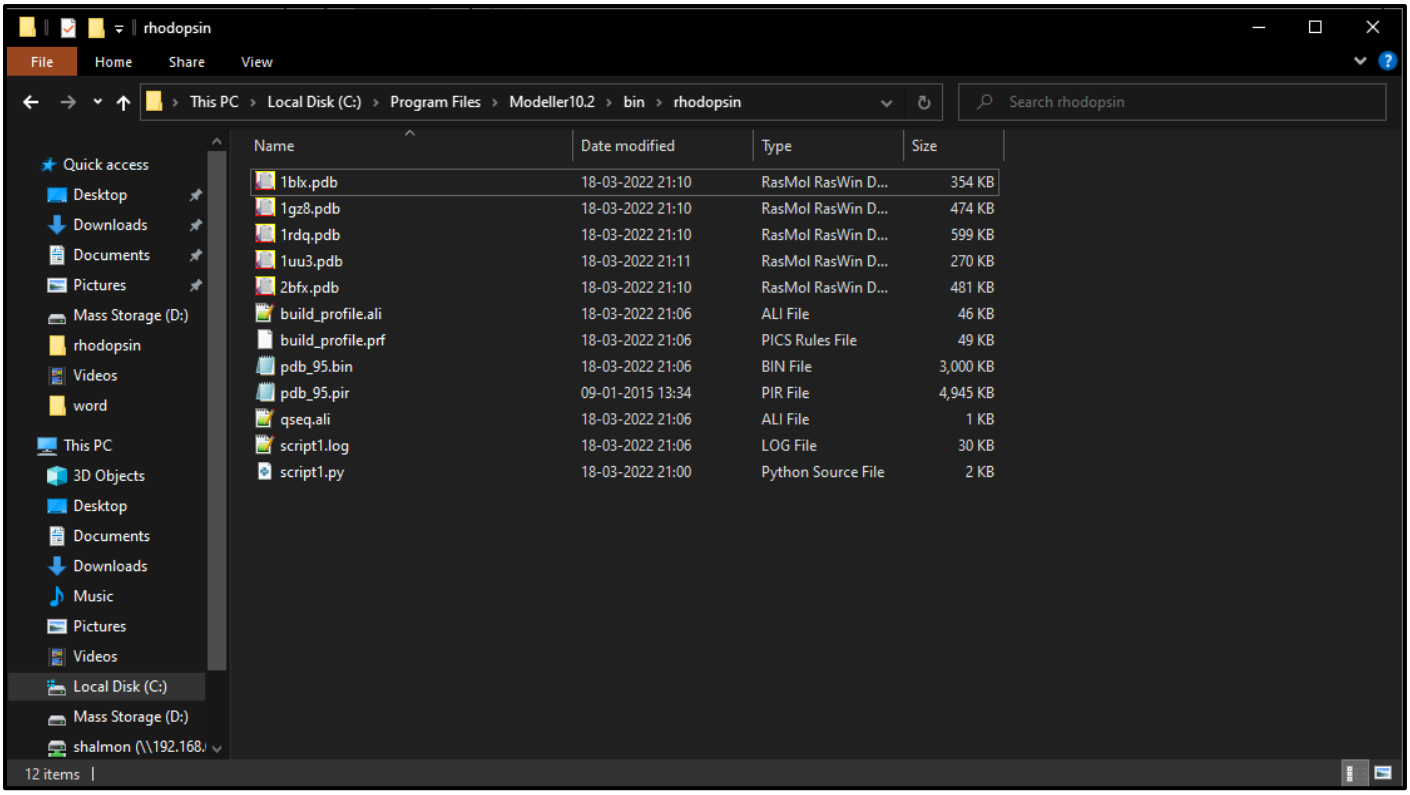


Fig12. Five structure download in PDB format

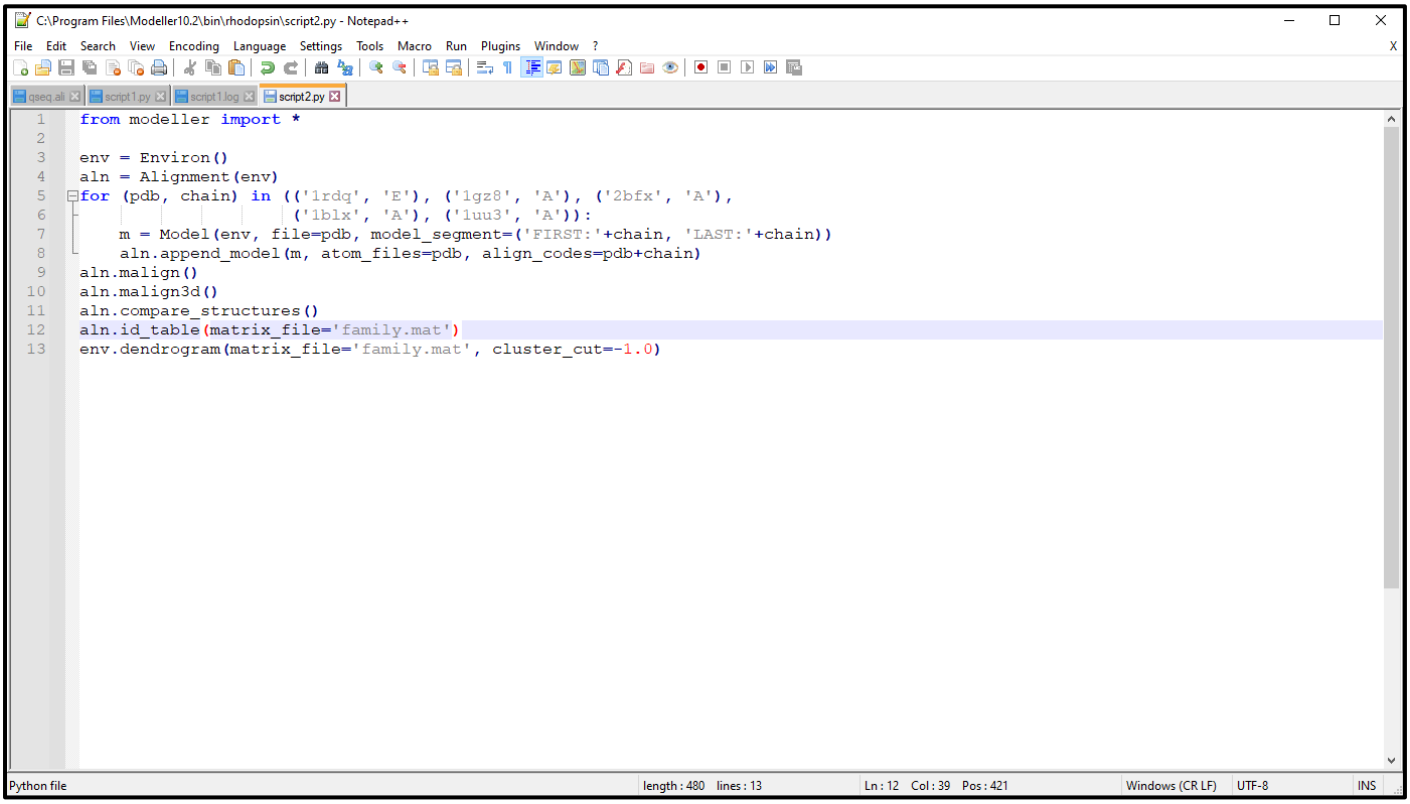


Fig13. Python script for selecting a template

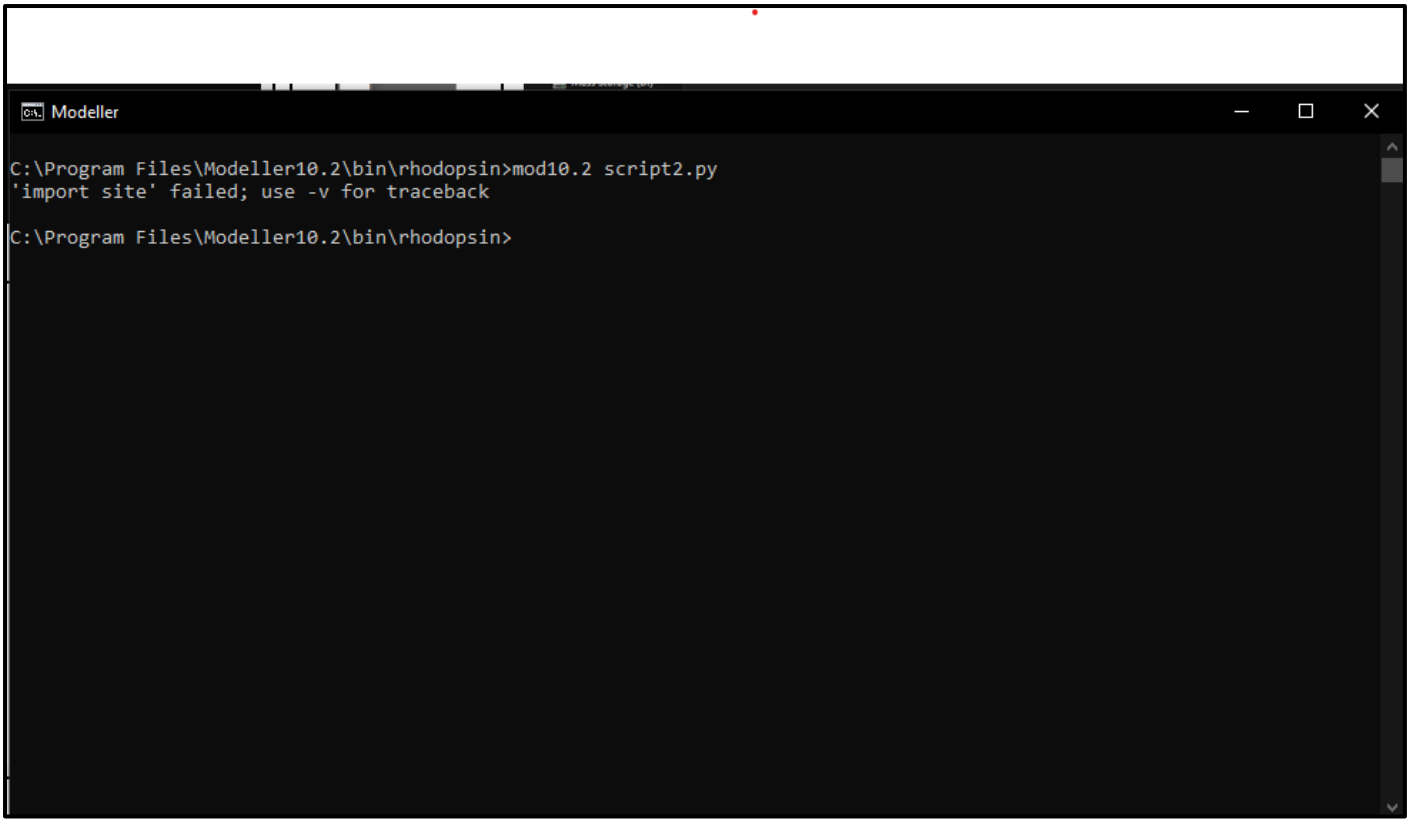


Fig14. Running script2.py

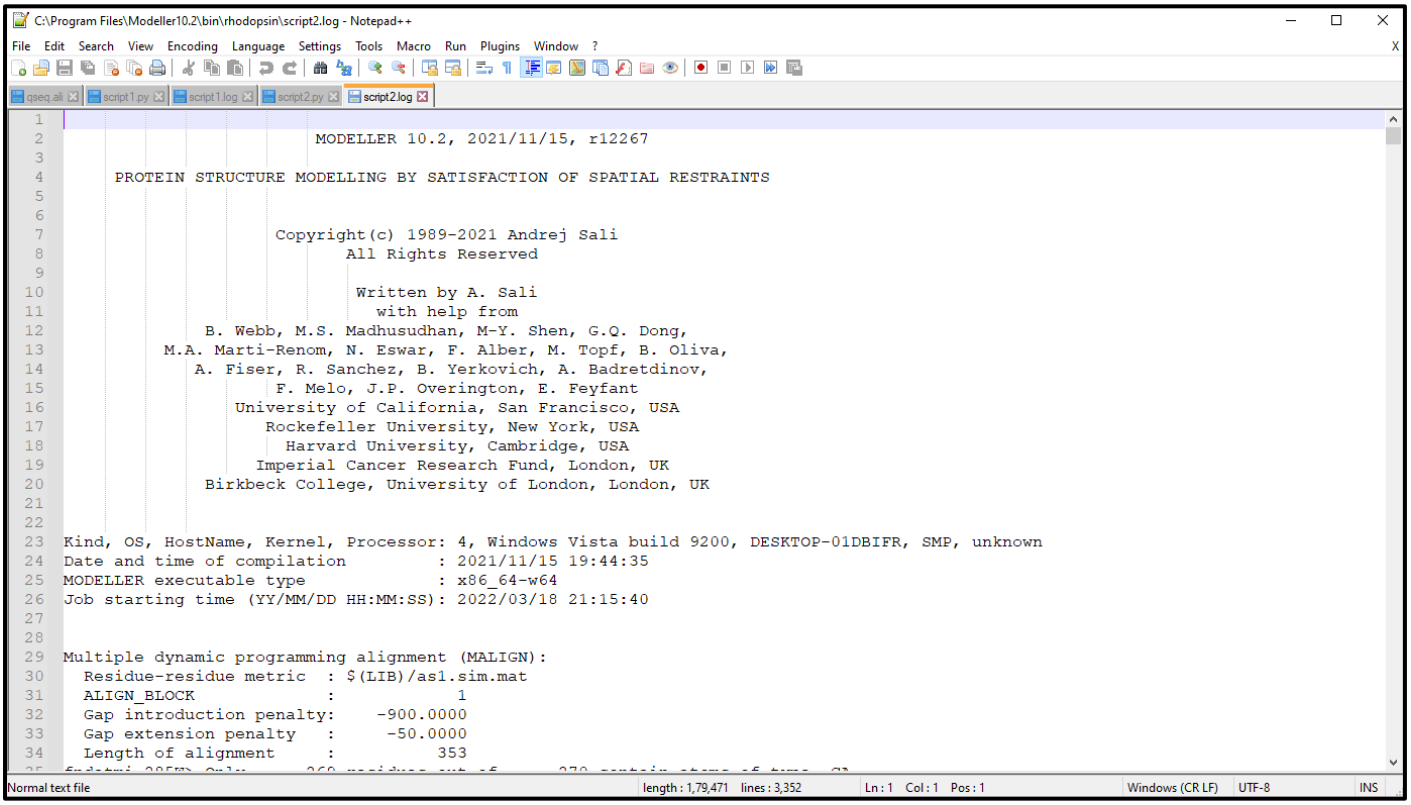


Fig15. Log file for script2

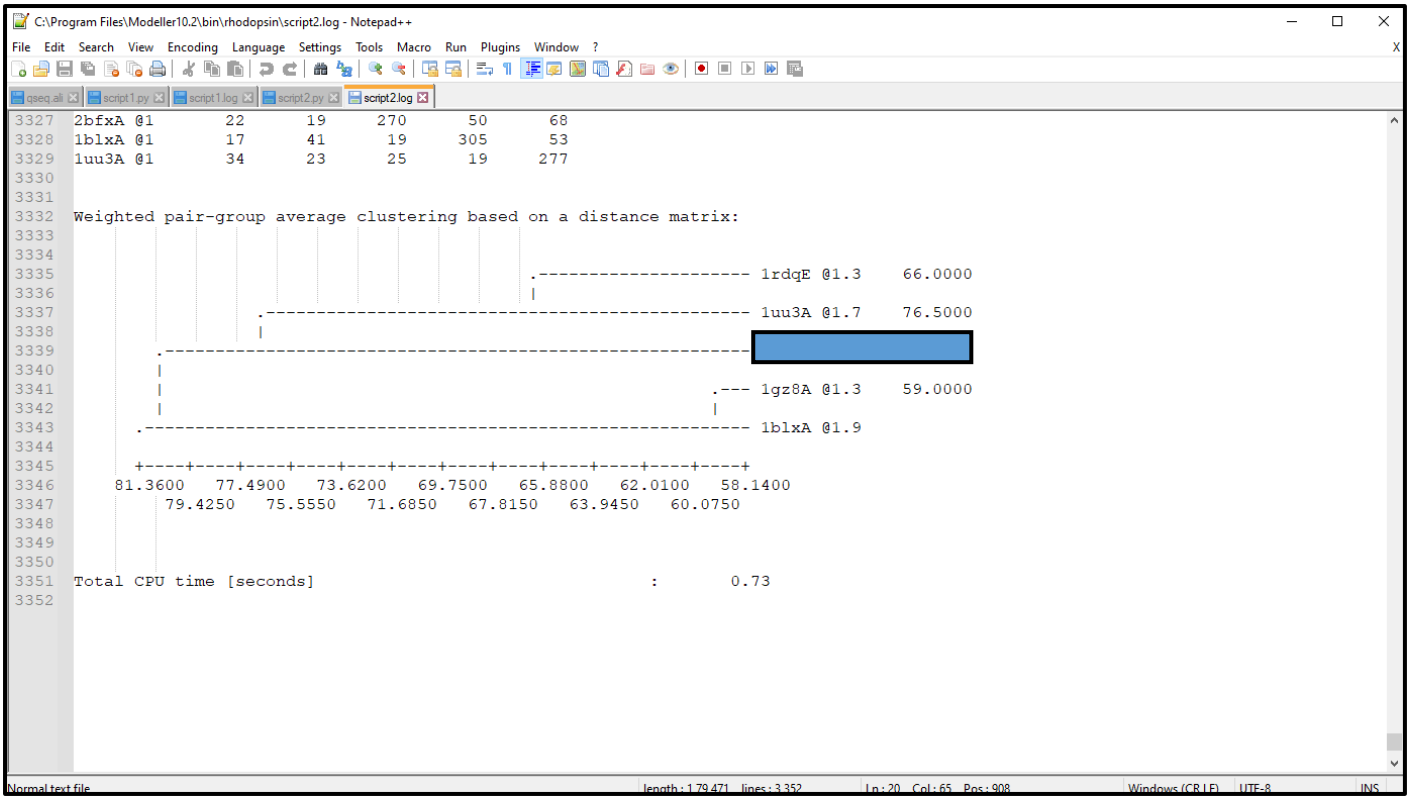


Fig15.1 Structure selected with low x=ray crystallography value and high NMR value

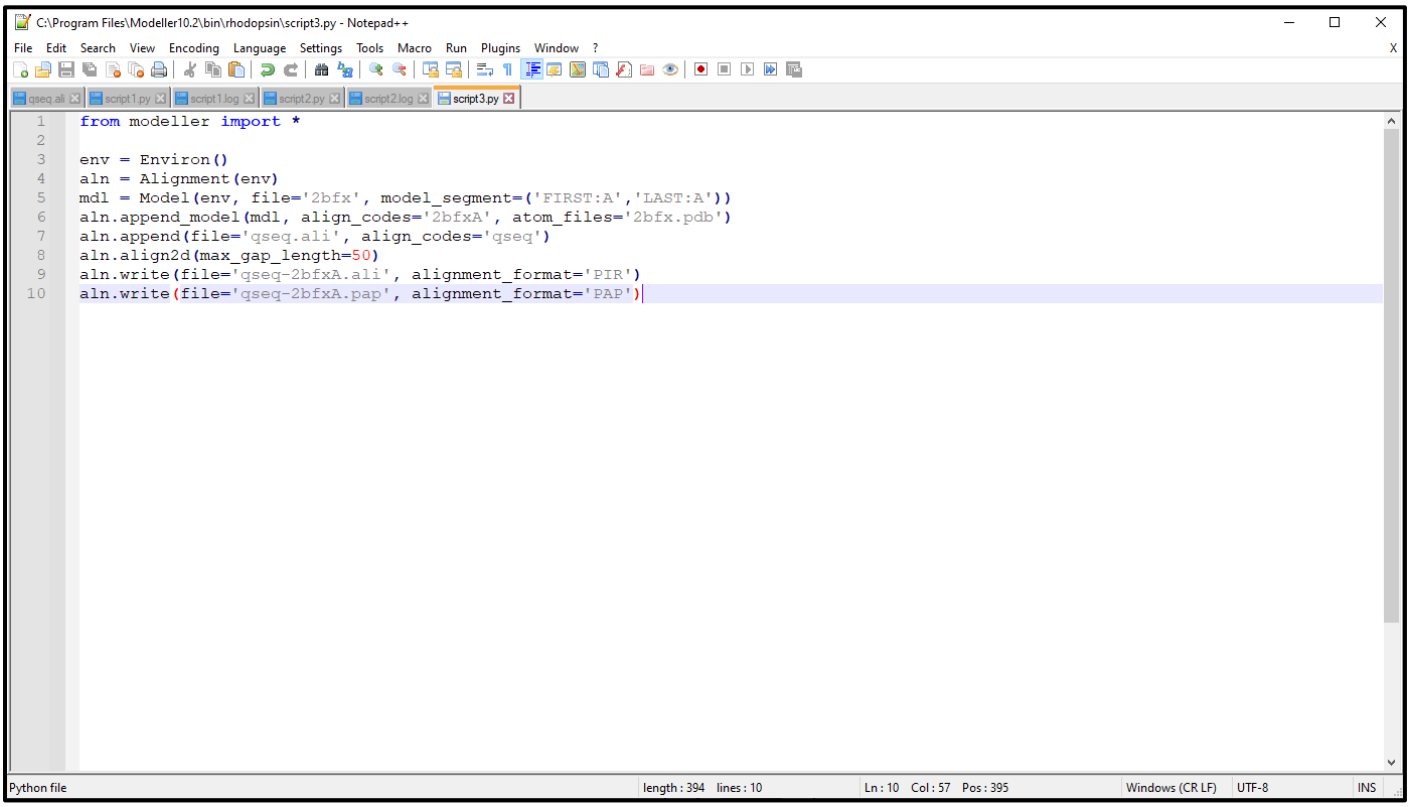


Fig16. Python script for aligning query with the template

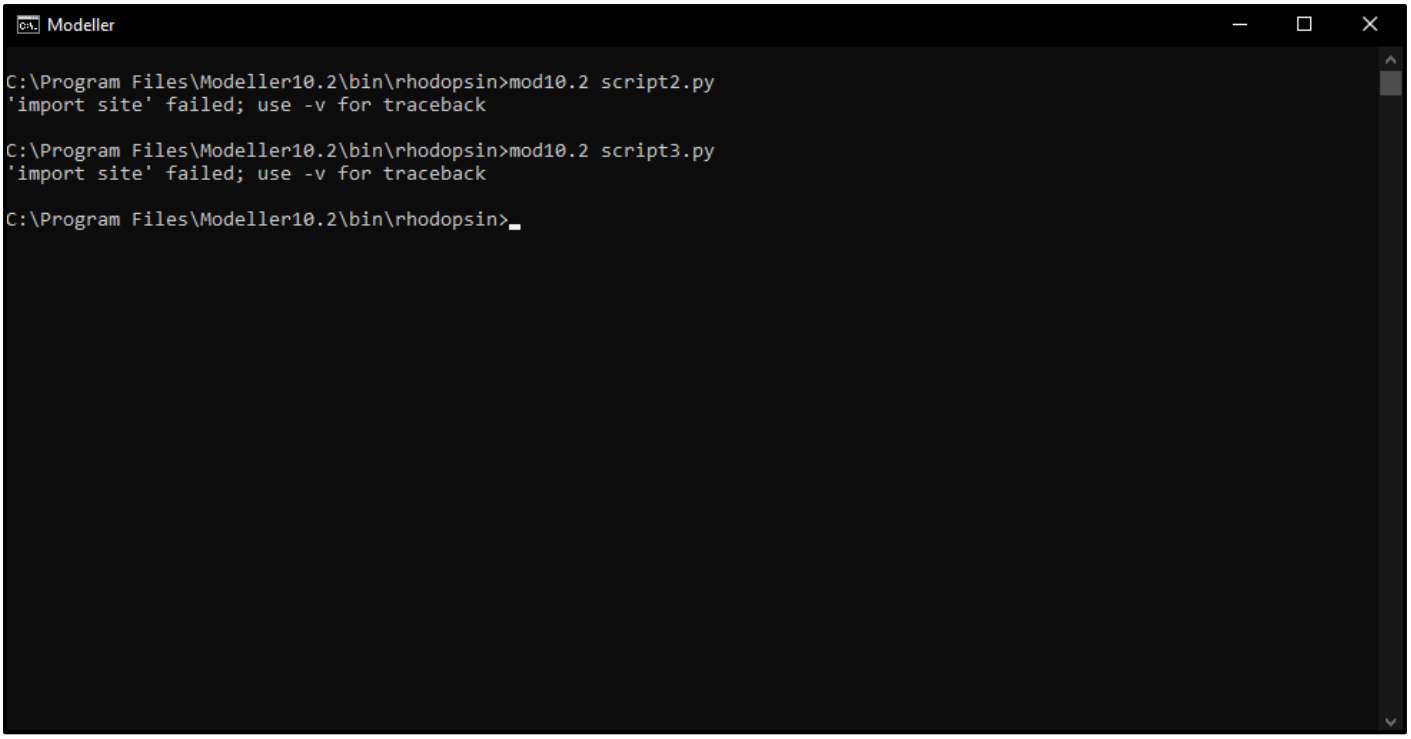


Fig17. Running script3.py

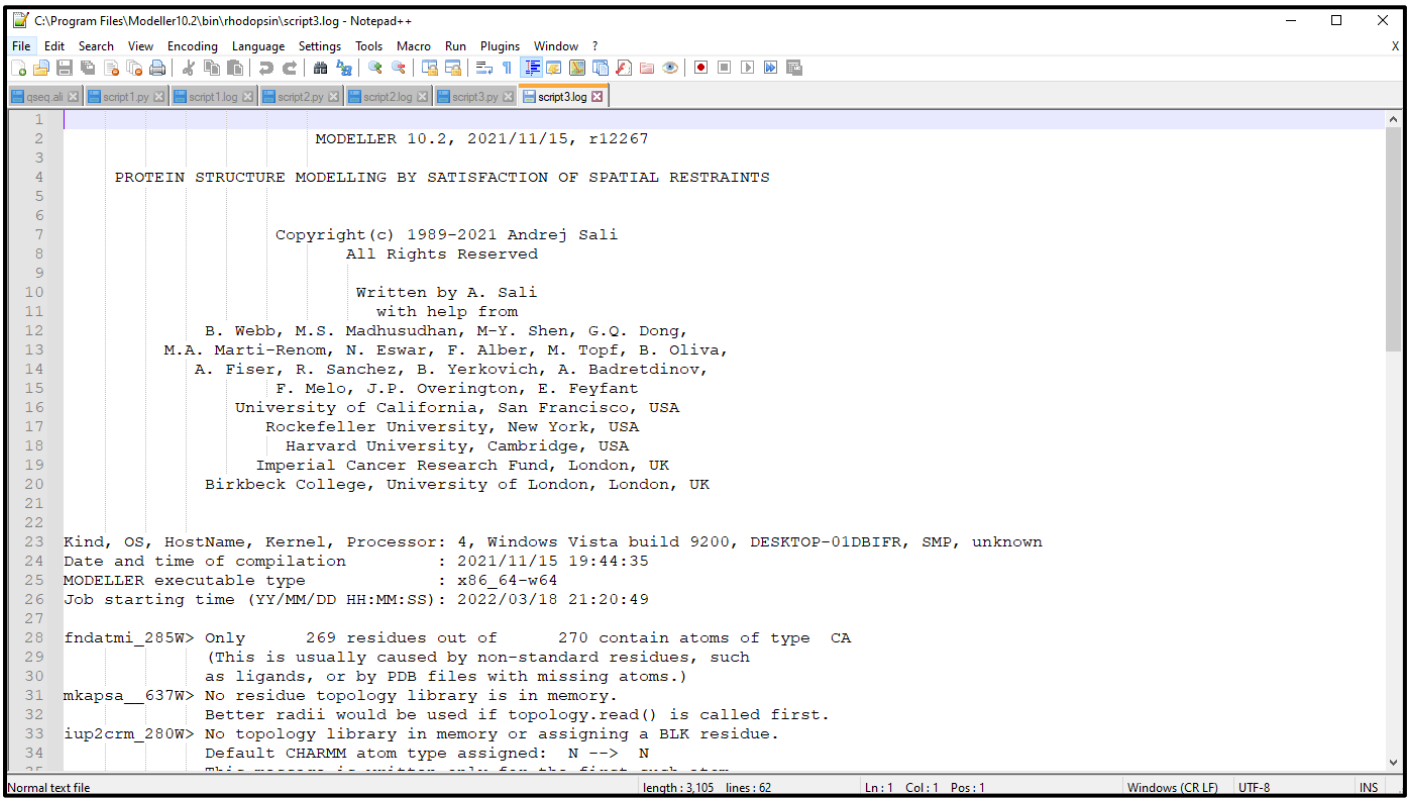


Fig18. Log file for script3

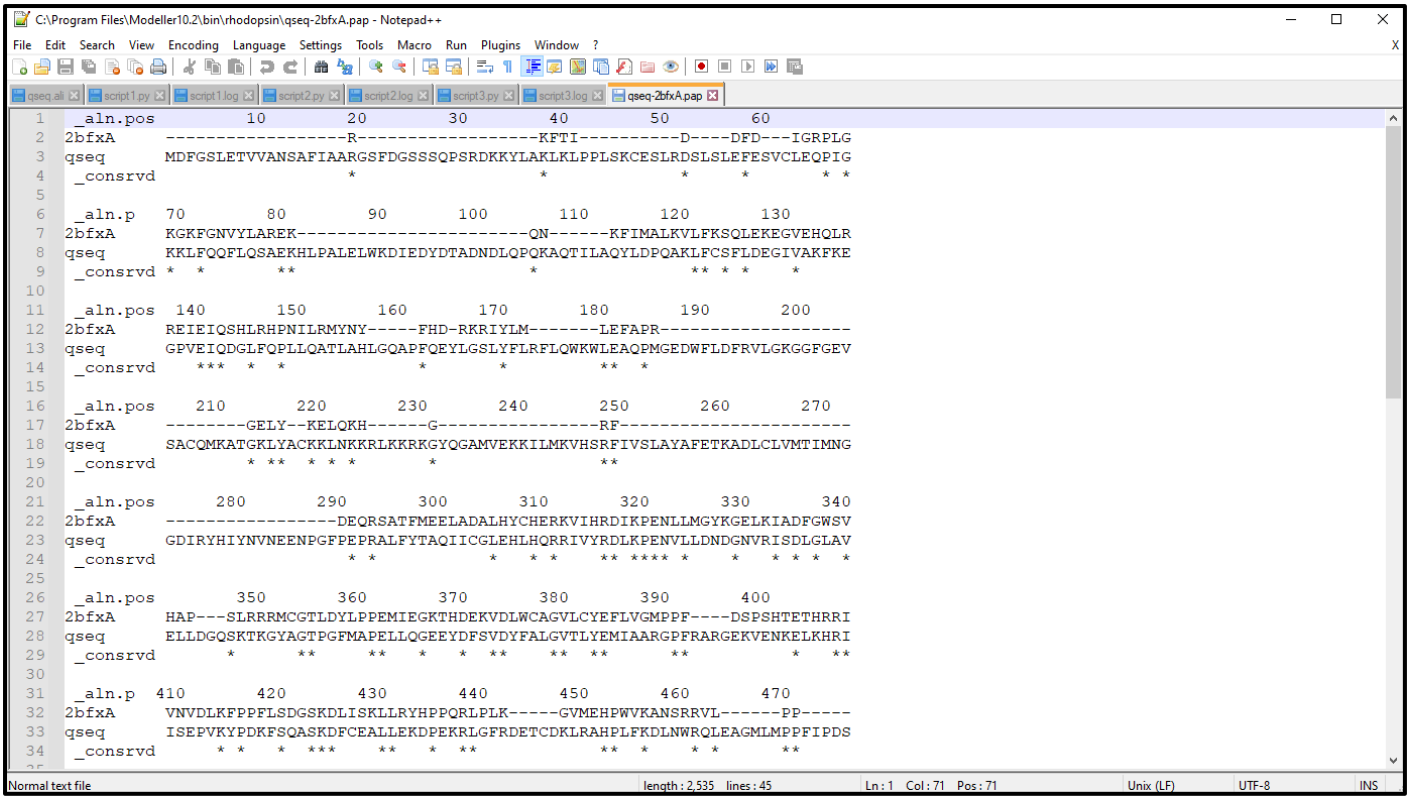


Fig19. Sequence alignment

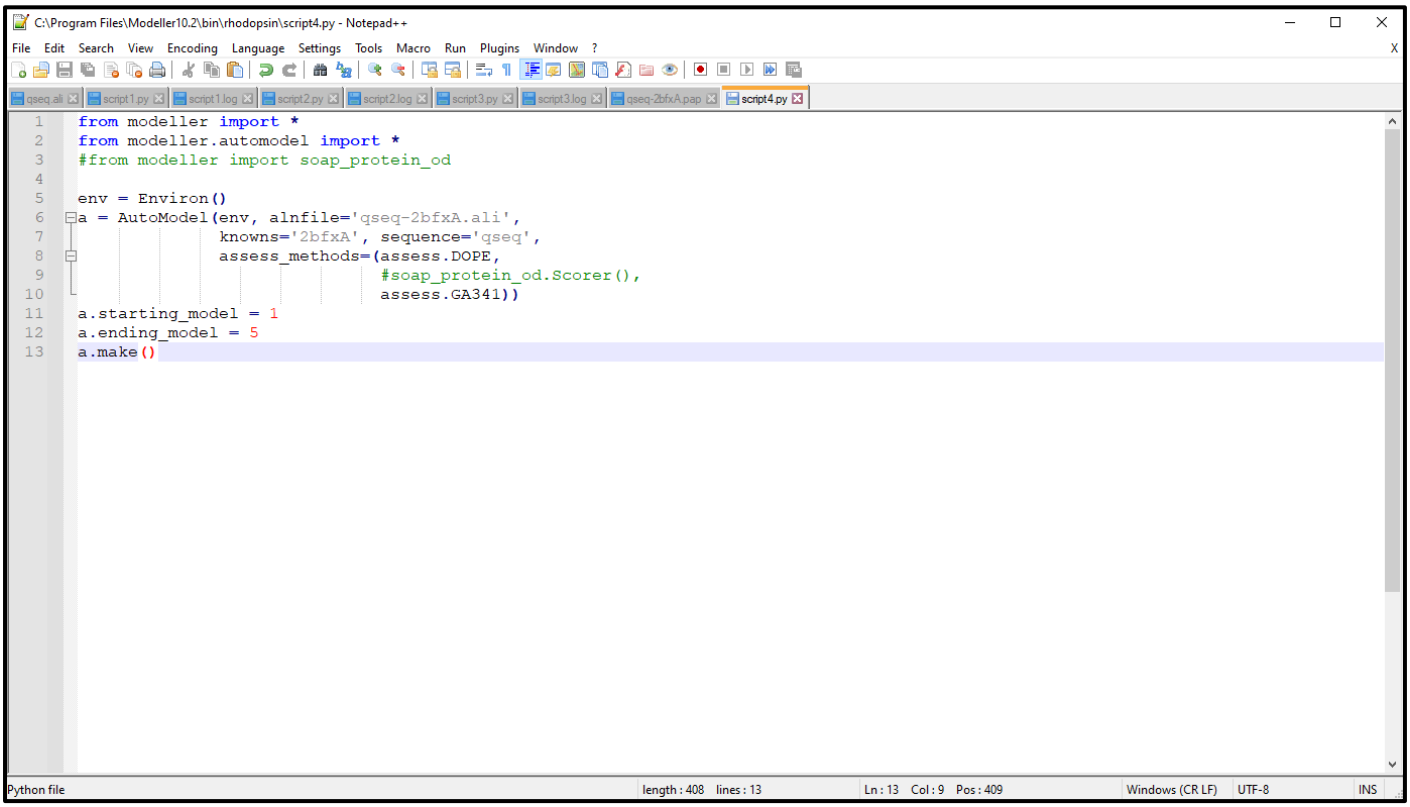


Fig20. Python script for model building

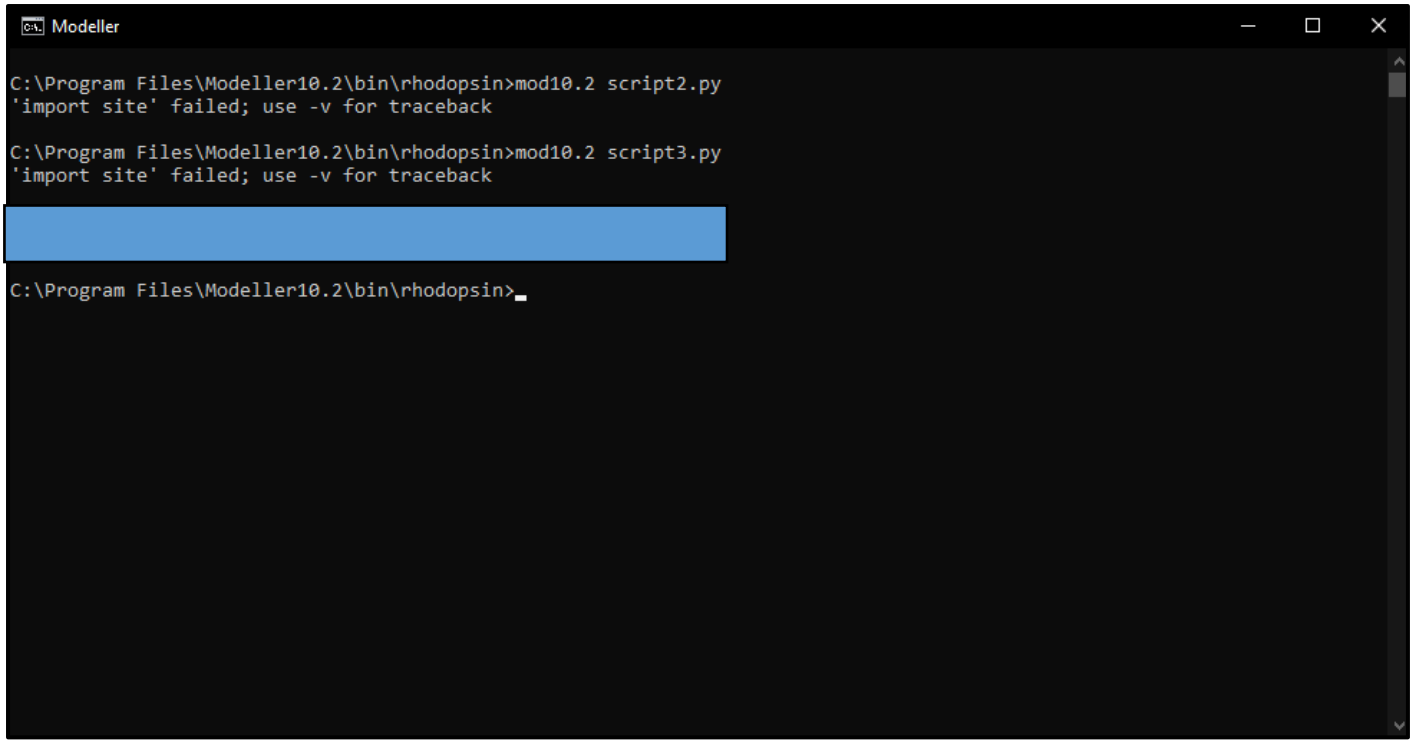


Fig21. Running script4.py

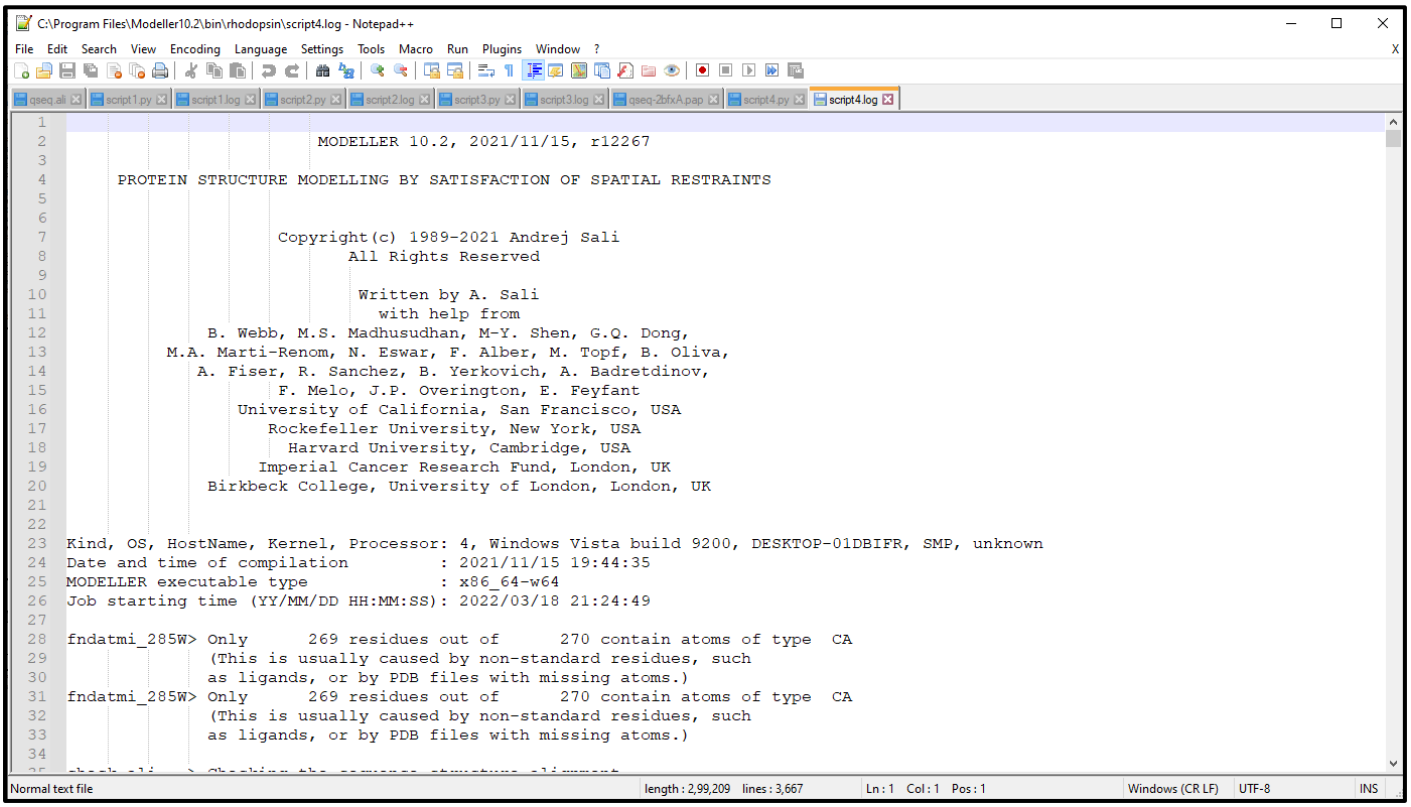


Fig22. Log file for script4

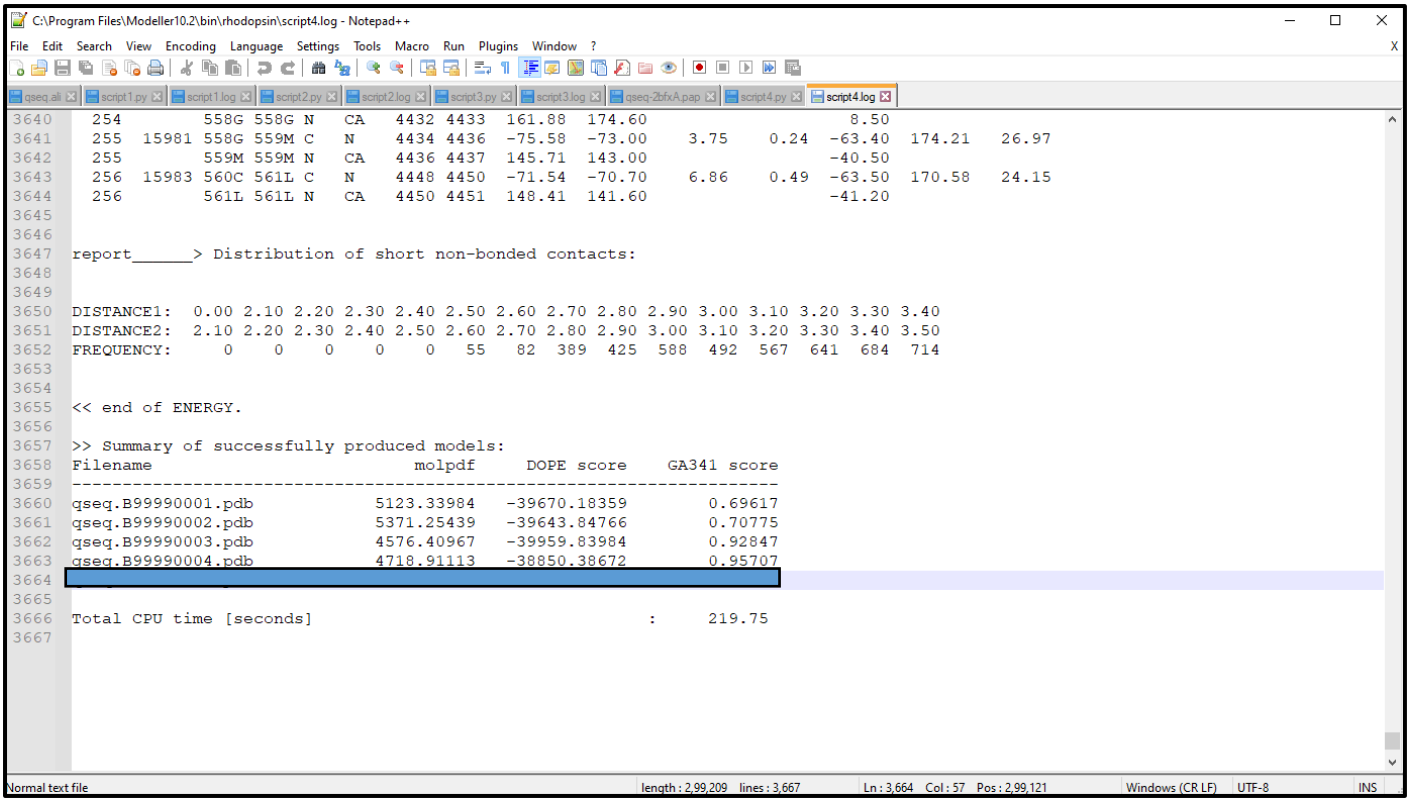


Fig23. Structure with lowest DOPE score selected as final model

RESULT:

Modeller was used to predict the tertiary structure of Rhodopsin

CONCLUSION:

Thus, modeller can be used to predict tertiary structures of proteins by comparative protein structure modelling. These tools give faster results than x-ray or NMR techniques and can be used by researchers to understand protein functions and drug designing.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 214-228.
2. Encyclopædia Britannica, inc. (n.d.). *Rhodopsin*. Encyclopædia Britannica. Retrieved March 18, 2022, from https://www.britannica.com/science/rhodopsin
3. Uniprot. (n.d.). Retrieved March 18, 2022, from https://www.uniprot.org/uniprot/Q15835.fasta

DATE: 18-03-22

WEBLEM 3b

I-TASSAR

(URL: https://zhanggroup.org/I-TASSER/)

AIM:

To perform tertiary structure prediction by threading approach using I-TASSER server for query rhodopsin.

INTRODUCTION:

Rhodopsin, also called visual purple, pigment-containing sensory protein that converts light into an electrical signal. Rhodopsin is found in a wide range of organisms, from vertebrates to bacteria. In many seeing animals, including humans, it is required for vision in dim light and is located in the retina of the eye—specifically, within the tightly packed disks that make up the outer segment of the retina's photoreceptive rod cells, which are specially adapted for vision under low-light conditions.

I-TASSER server is an on-line platform that implements the I-TASSER based algorithms for protein structure and function predictions. It allows academic users to automatically generate high-quality model predictions of 3D structure and biological function of protein molecules from their amino acid sequences.

METHODOLOGY:

1. Open homepage for I-TASSER. (URL: https://zhanggroup.org/I-TASSER/)
2. Complete registration.
3. Submit FASTA sequence for kinase.
4. Observe and interpret results.

OBSERVATION:

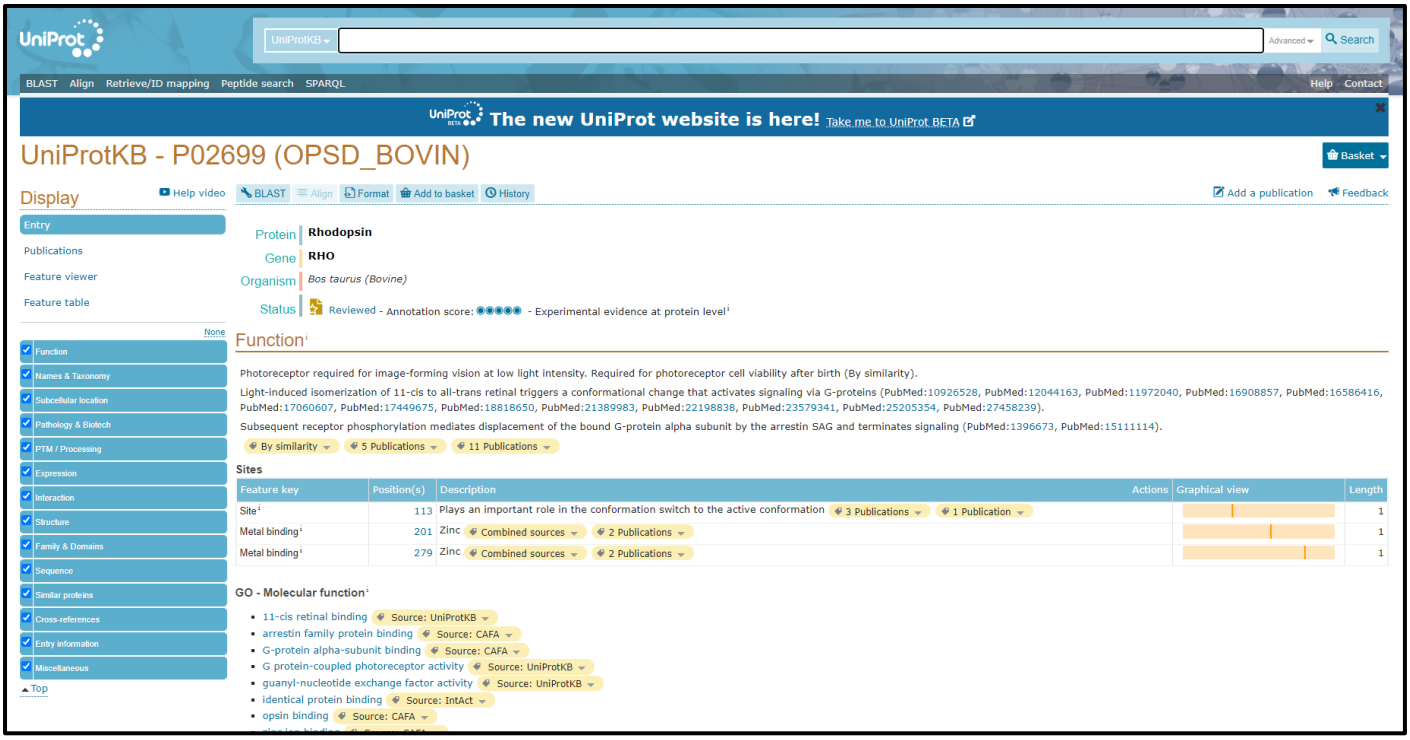


Fig1. Result page for Rhodopsin in UniProt database

```
>sp|P02699|OPSD_BOVIN Rhodopsin OS=Bos taurus OX=9913 GN=RHO PE=1 SV=1
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
EGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAV
YNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA
```

Fig2. FASTA sequence for Rhodopsin

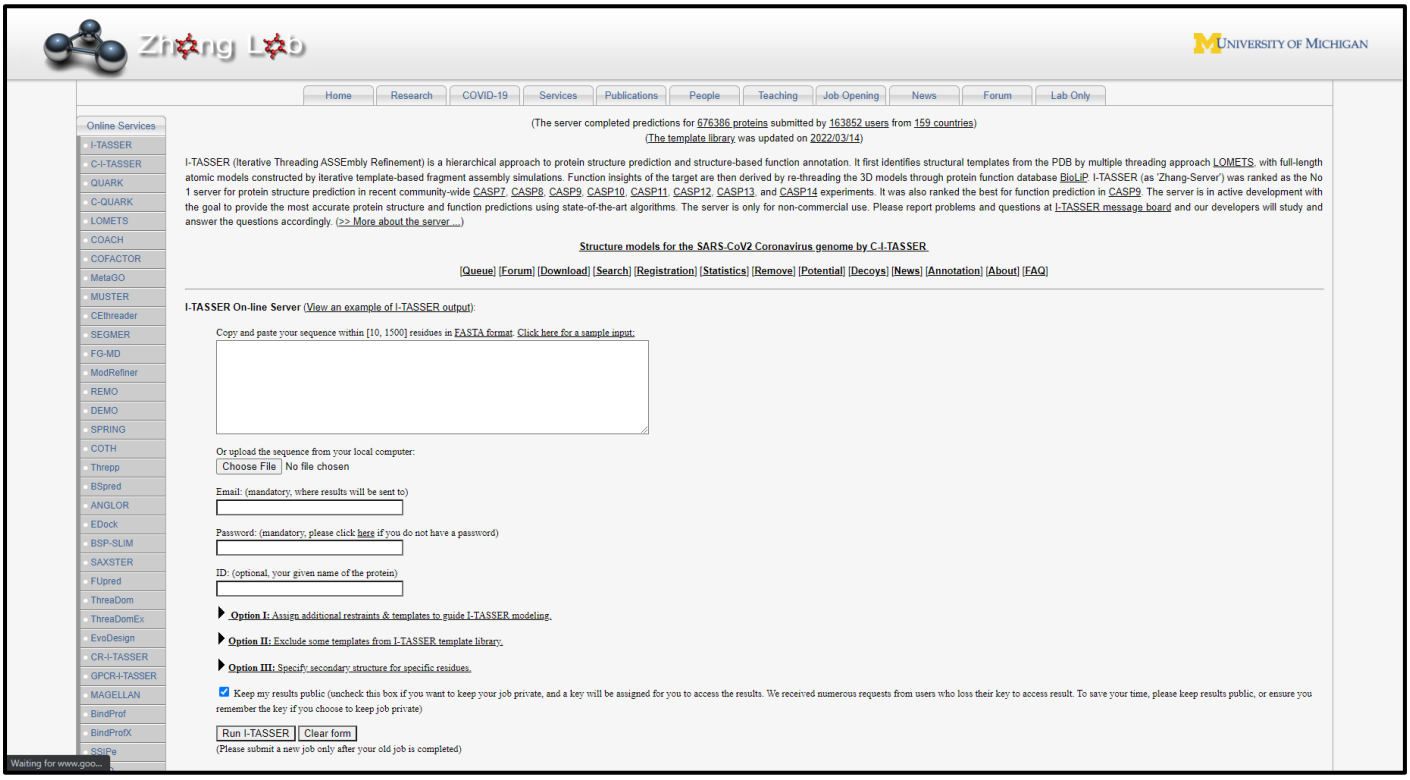


Fig3. Homepage for I-TASSAR

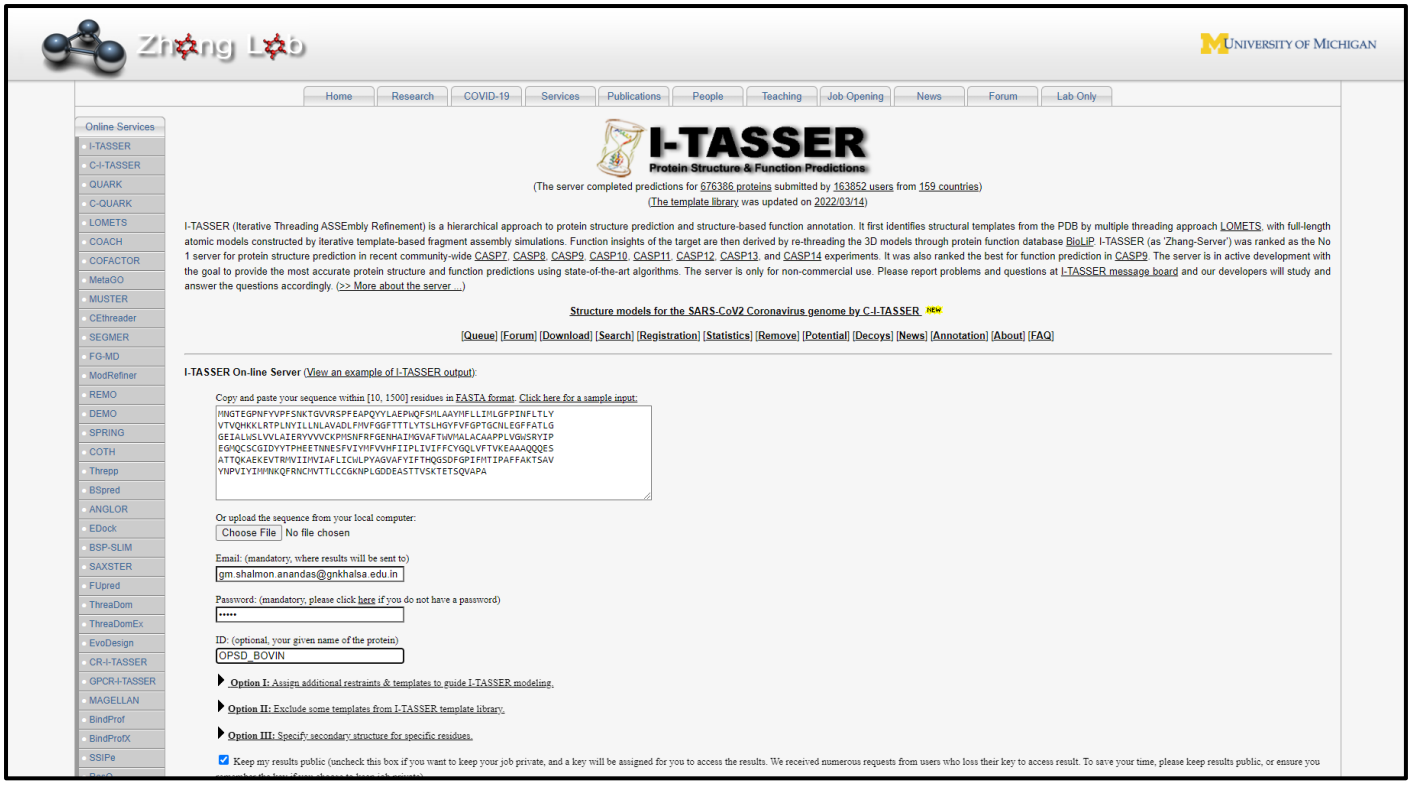


Fig4. Submission of query

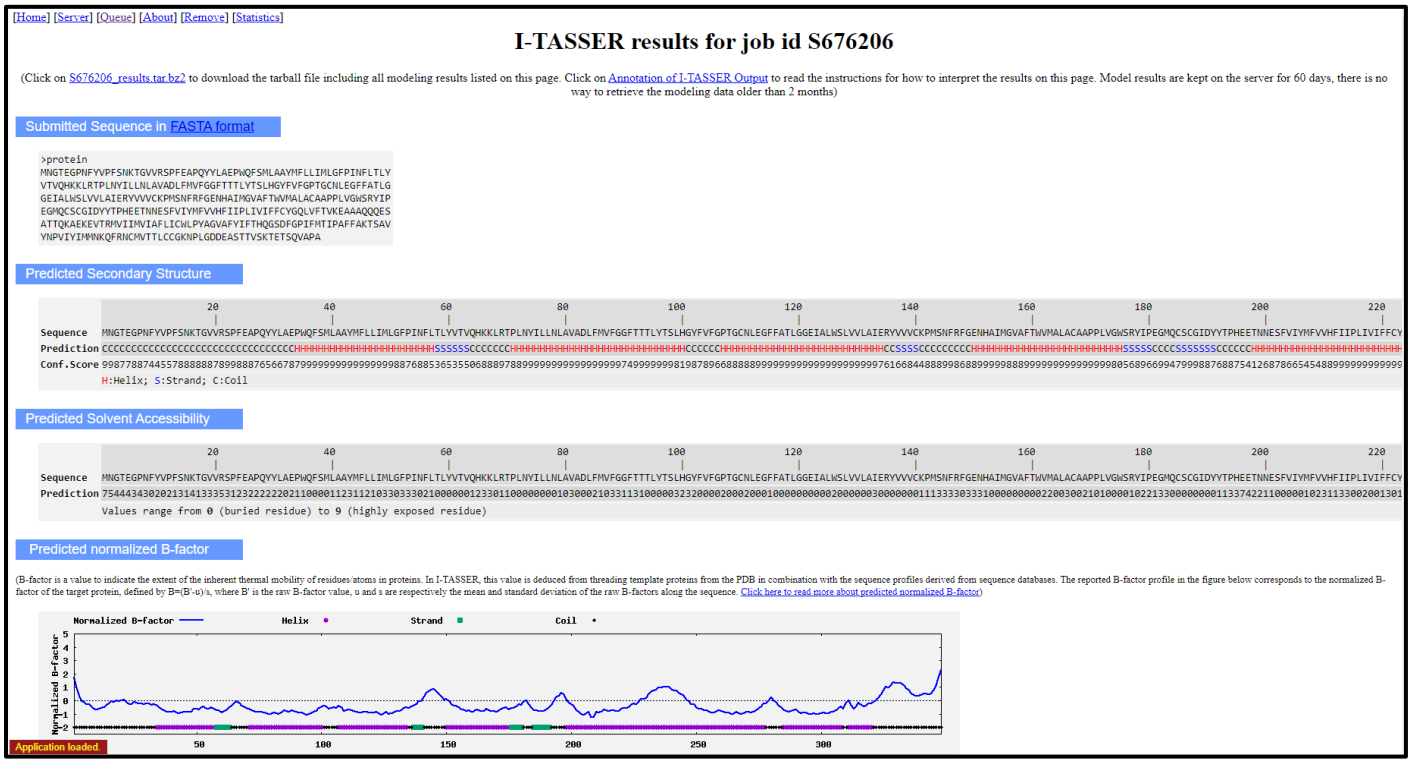


Fig5. Result for predicted secondary structure and normalized B-factor

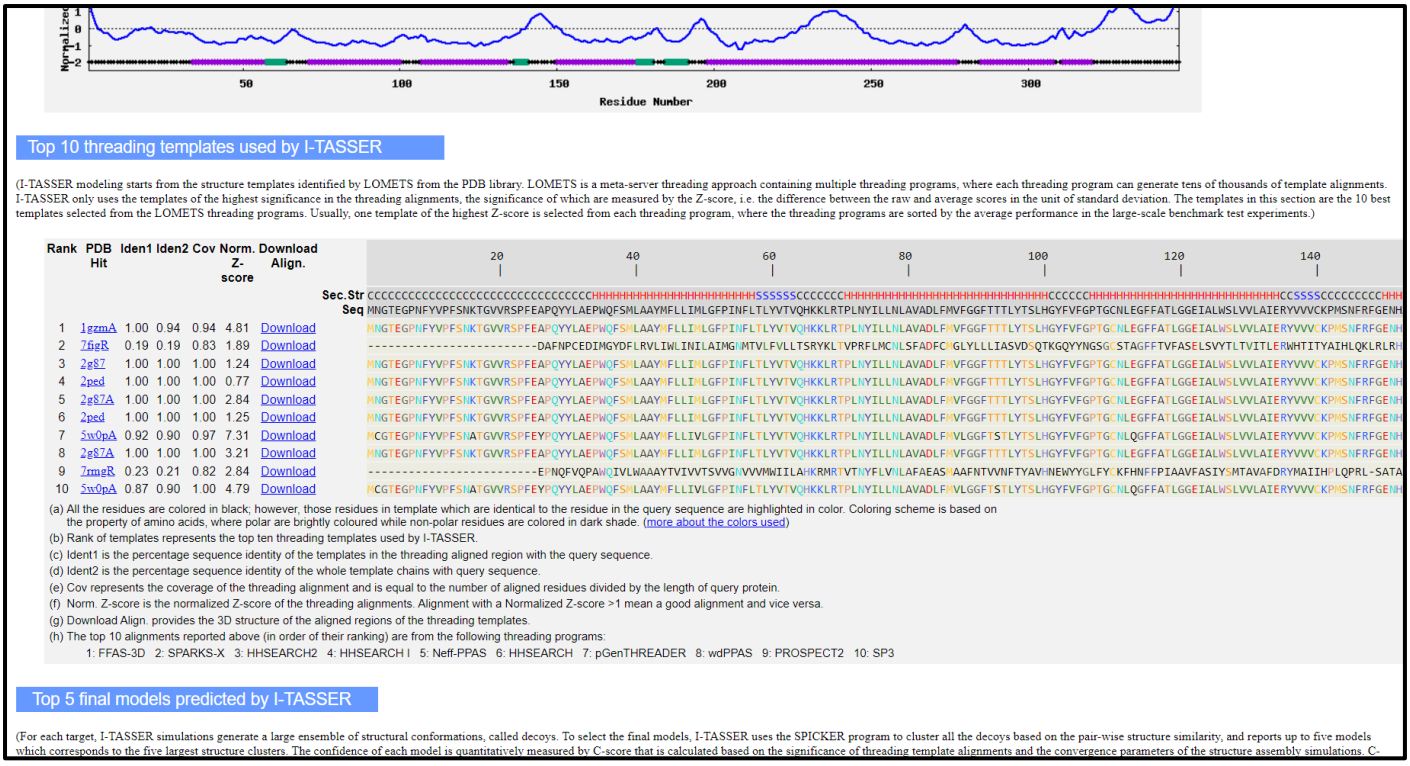


Fig7. Result for top 10 threading templates

a. All the residues are coloured in black; however, those residues in template which are identical to the residue in the query sequence are highlighted in colour. Colouring scheme is based on the property of amino acids, where polar are brightly coloured while non-polar residues are coloured in dark shade. (More about the colours used)
b. Rank of templates represents the top ten threading templates used by I-TASSER.
c. Iden1 is the percentage sequence identity of the templates in the threading aligned region with the query sequence.
d. Iden2 is the percentage sequence identity of the whole template chains with query sequence.
e. Cov represents the coverage of the threading alignment and is equal to the number of aligned residues divided by the length of query protein.
f. Norm. Z-score is the normalized Z-score of the threading alignments. Alignment with a Normalized Z-score >1 mean a good alignment and vice versa.
g. Download Align. provides the 3D structure of the aligned regions of the threading templates.
h. The top 10 alignments reported above (in order of their ranking) are from the following threading programs: 1: FFAS-3D 2: SPARKS-X 3: HHSEARCH2 4: HHSEARCH I 5: Neff-PPAS 6: HHSEARCH 7: pGenTHREADER 8: wdPPAS 9: PROSPECT2 10: SP3

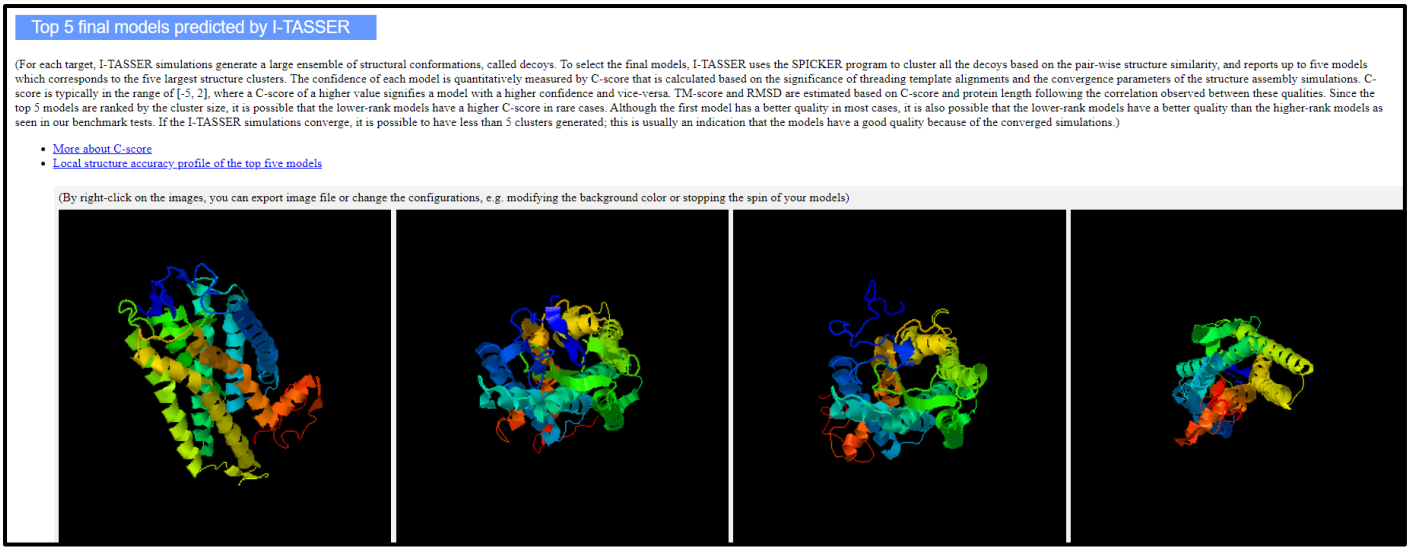


Fig8. Result for top 5 final models predicted

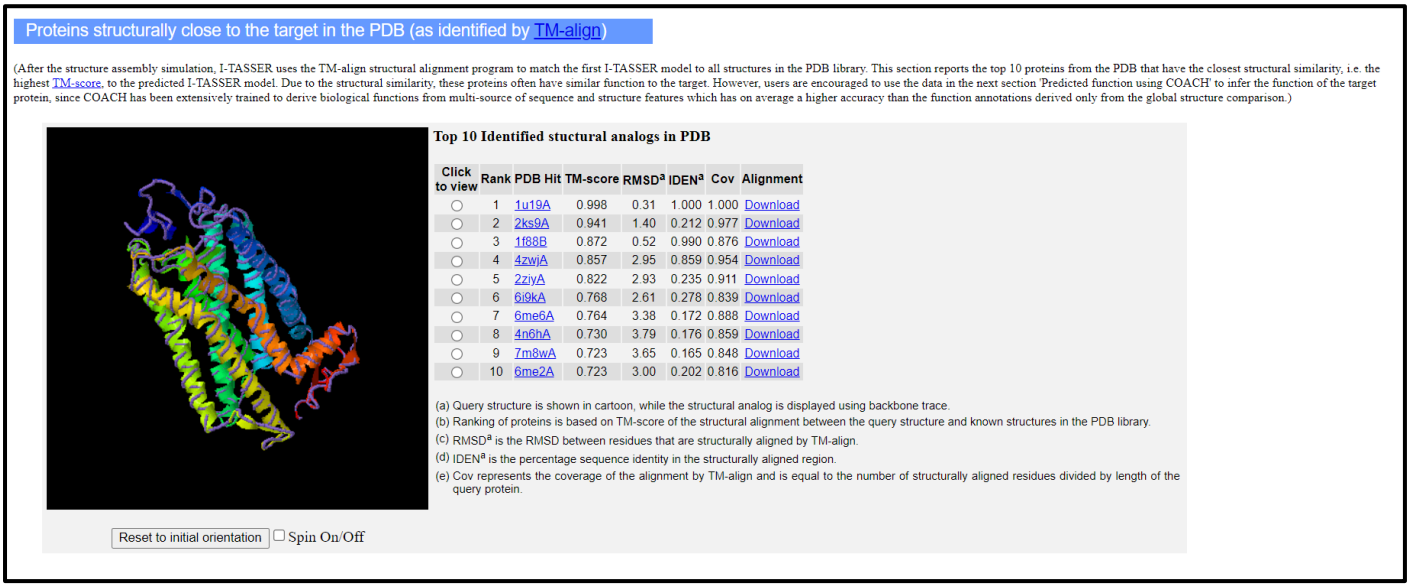


Fig9. Result for proteins that are structurally close to target

a. Query structure is shown in cartoon, while the structural analog is displayed using backbone trace.
b. Ranking of proteins is based on TM-score of the structural alignment between the query structure and known structures in the PDB library
c. RMSDa is the RMSD between residues that are structurally aligned by TM-align
d. IDENa is the percentage sequence identity in the structurally aligned region.
e. Cov represents the coverage of the alignment by TM-align and is equal to the number of structurally aligned residues divided by length of the query protein.

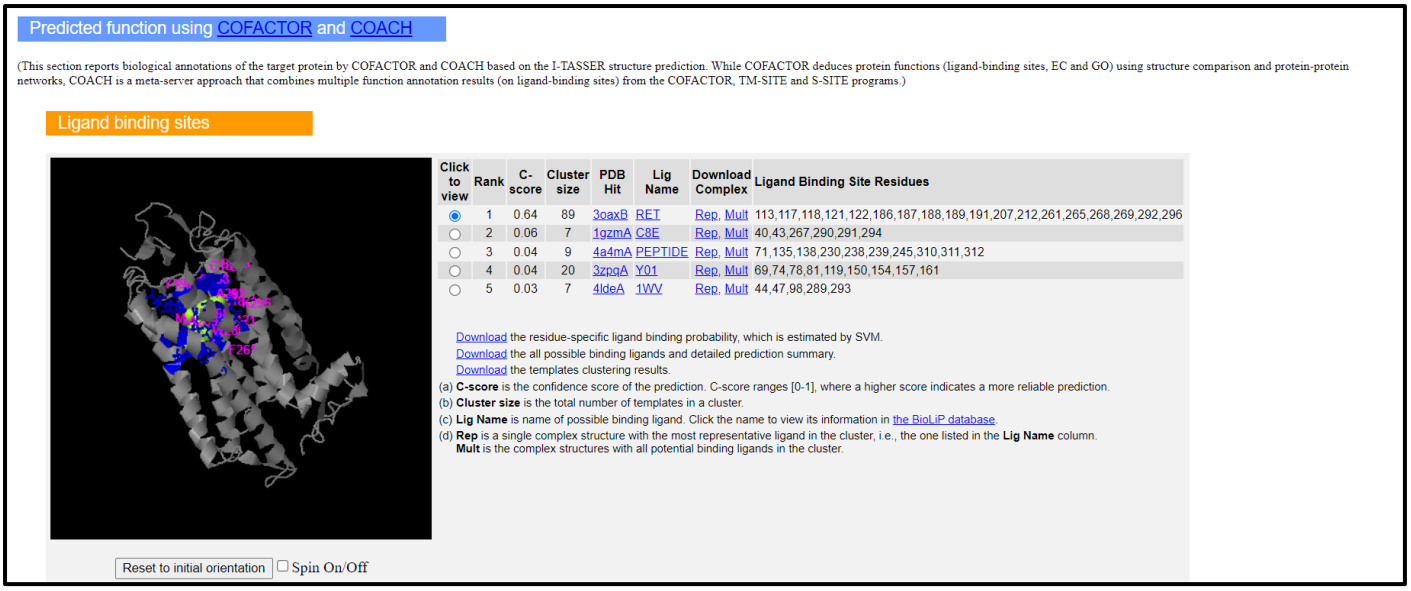


Fig10. Result for predicted fuctions

a. C-score is the confidence score of the prediction. C-scores ranges [0-1], where a higher score indicated a more reliable prediction.
b. Cluster size is the total number of templates in a cluster.
c. Lig Name is name of possible binding ligand. Click the name to view its information in the BioLiP database.
d. Rep is a single complex structure with the most representative ligand in the cluster, i.e., the one listed in the Lig Name column. Mult is the complex structures with all potential binding ligands in the cluster.

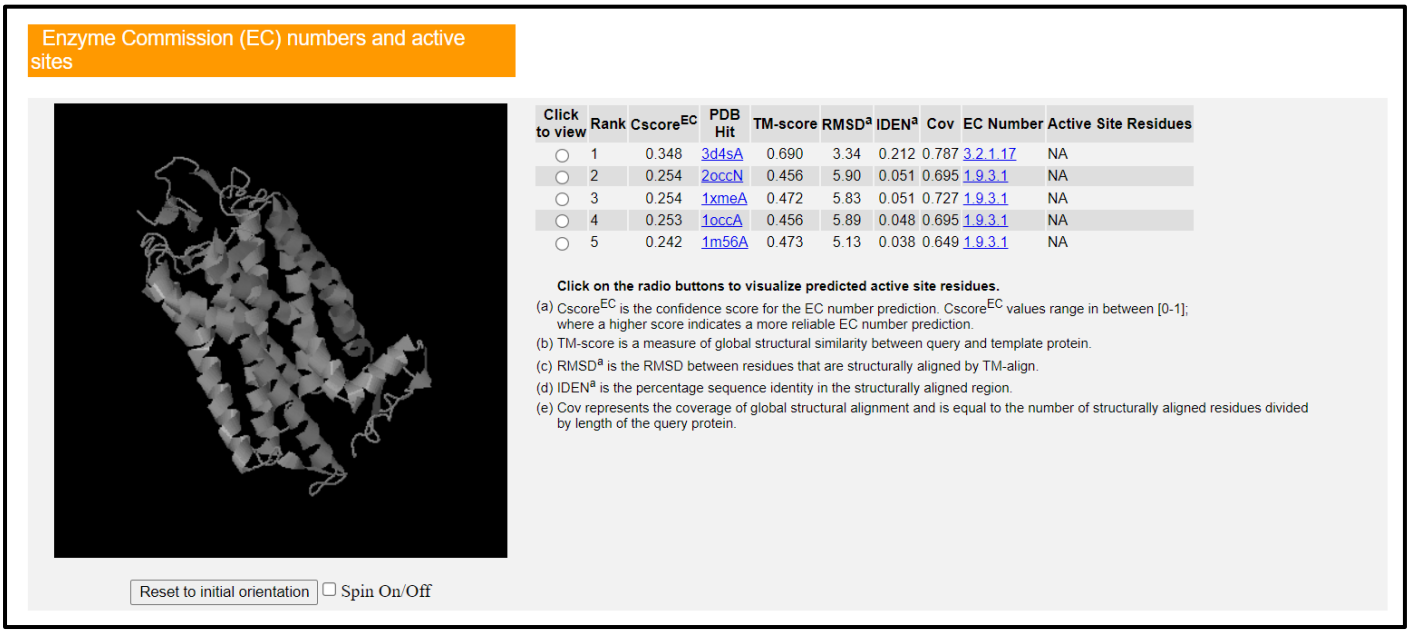
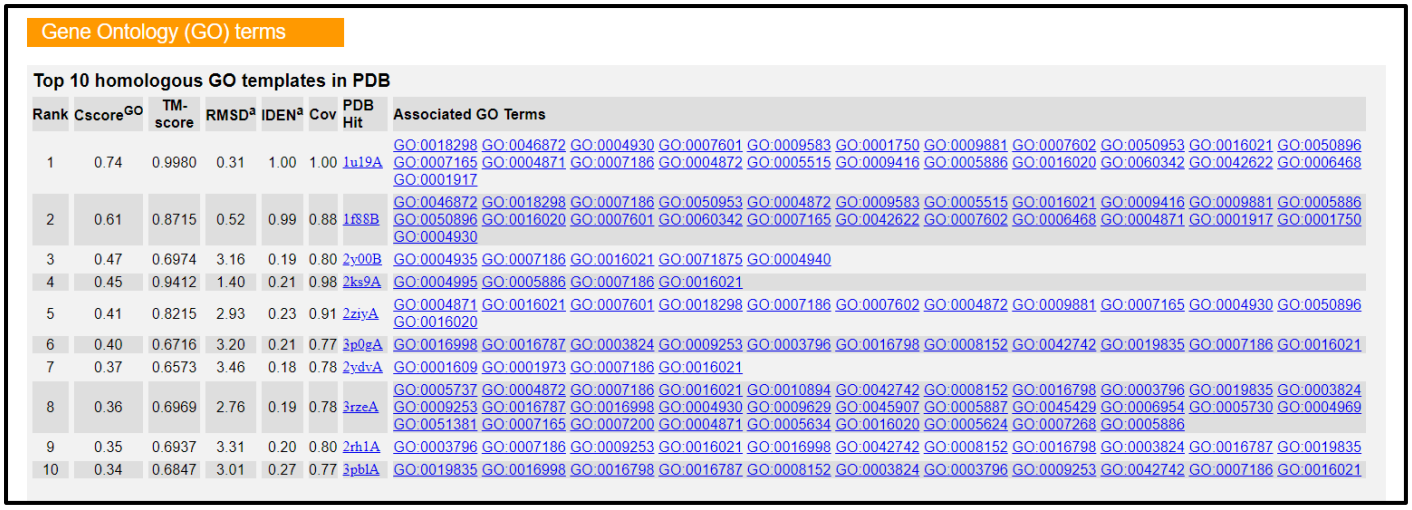


Fig11. Enzyme commission (EC) numbers and active sites

a. CscoreEC is the confidence score for the EC number prediction. CscoreEC values range in between [0-1]; where a higher score indicates a more reliable EC number prediction.
b. TM-score is a measure of global structural similarity between query and template protein.
c. RMSDa is the RMSD between residues that are structurally aligned by TM-align.
d. IDENa is the percentage sequence identity in the structurally aligned region.
e. Cov represents the coverage of global structural alignment and is equal to the number of structurally aligned residues divided by length of the query protein.



Gene Ontology (GO) terms

Top 10 homologous GO templates in PDB

Rank	Cscore^{GO}	TM-score	RMSD^{a}	IDEN^{a}	Cov	PDB Hit	Associated GO Terms
1	0.74	0.9980	0.31	1.00	1.00	1u19A	GO:0018298 GO:0046872 GO:0004930 GO:0007601 GO:0009583 GO:0001750 GO:0009881 GO:0007602 GO:0050953 GO:0016021 GO:0050896 GO:0007165 GO:0004871 GO:0007186 GO:0004872 GO:0005515 GO:0009416 GO:0005886 GO:0016020 GO:0060342 GO:0042622 GO:0006468 GO:0001917
2	0.61	0.8715	0.52	0.99	0.88	1f88B	GO:0046872 GO:0018298 GO:0007186 GO:0050953 GO:0004872 GO:0009583 GO:0005515 GO:0016021 GO:0009416 GO:0009881 GO:0005886 GO:0050896 GO:0016020 GO:0007601 GO:0060342 GO:0007165 GO:0042622 GO:0007602 GO:0006468 GO:0004871 GO:0001917 GO:0001750 GO:0004930
3	0.47	0.6974	3.16	0.19	0.80	2y00B	GO:0004935 GO:0007186 GO:0016021 GO:0071875 GO:0004940
4	0.45	0.9412	1.40	0.21	0.98	2ks9A	GO:0004995 GO:0005886 GO:0007186 GO:0016021
5	0.41	0.8215	2.93	0.23	0.91	2ziyA	GO:0004871 GO:0016021 GO:0007601 GO:0018298 GO:0007186 GO:0007602 GO:0004872 GO:0009881 GO:0007165 GO:0004930 GO:0050896 GO:0016020
6	0.40	0.6716	3.20	0.21	0.77	3p0gA	GO:0016998 GO:0016787 GO:0003824 GO:0009253 GO:0003796 GO:0016798 GO:0008152 GO:0042742 GO:0019835 GO:0007186 GO:0016021
7	0.37	0.6573	3.46	0.18	0.78	2ydvA	GO:0001609 GO:0001973 GO:0007186 GO:0016021
8	0.36	0.6969	2.76	0.19	0.78	3rzeA	GO:0005737 GO:0004872 GO:0007186 GO:0016021 GO:0010894 GO:0042742 GO:0008152 GO:0016798 GO:0003796 GO:0019835 GO:0003824 GO:0009253 GO:0016787 GO:0016998 GO:0004930 GO:0009629 GO:0045907 GO:0005887 GO:0045429 GO:0006954 GO:0005730 GO:0004969 GO:0051381 GO:0007165 GO:0007200 GO:0004871 GO:0005634 GO:0016020 GO:0005624 GO:0007268 GO:0005886
9	0.35	0.6937	3.31	0.20	0.80	2rh1A	GO:0003796 GO:0007186 GO:0009253 GO:0016021 GO:0016998 GO:0042742 GO:0008152 GO:0016798 GO:0003824 GO:0016787 GO:0019835
10	0.34	0.6847	3.01	0.27	0.77	3pblA	GO:0019835 GO:0016998 GO:0016798 GO:0016787 GO:0008152 GO:0003824 GO:0003796 GO:0009253 GO:0042742 GO:0007186 GO:0016021

Fig12. Gene Ontology (GO) terms

a. CscoreGO is a combined measure for evaluating global and local similarity between query and template protein. It's range is [0-1] and higher values indicate more confident predictions.
b. TM-score is a measure of global structural similarity between query and template protein.
c. RMSDa is the RMSD between residues that are structurally aligned by TM-align.
d. IDENa is the percentage sequence identity in the structurally aligned region.
e. Cov represents the coverage of global structural alignment and is equal to the number of structurally aligned residues divided by length of the query protein.
f. The second table shows a consensus GO terms amongst the top scoring templates. The GO-Score associated with each prediction is defined as the average weight of the GO term, where the weights are assigned based on CscoreGO of the template.

RESULT:

I-TASSER was used to predict the tertiary structure of Kinase based on threading approach. The information regarding solvent accessibility, normalized B-factor, top 10 threading templates, top 5 final models, proteins that are structurally close to target, functions and active sites were predicted.

CONCLUSION:

Thus, I-TASSER can be used to predict tertiary structures of proteins by threading method. These tools give faster results than x-ray or NMR techniques and can be used by researchers to understand protein functions and drug designing.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 214-228.
2. kinase | Definition, Biology, & Function. (n.d.). Encyclopedia Britannica. Retrieved March 8, 2022, from https://www.britannica.com/science/kinase
3. I-TASSER server for protein structure and function prediction. (n.d.-b). Zhanggroup.org. Retrieved March 8, 2022, from https://zhanggroup.org/I-TASSER/
4. I-TASSER results. (n.d.). Zhanggroup.org. Retrieved March 8, 2022, from https://zhanggroup.org//I-TASSER/output/S673761/

DATE: 18-03-22

WEBLEM 3c

ROBETTA

(URL: https://robetta.bakerlab.org/)

AIM:

To perform tertiary structure prediction by Ab-Initio approach using ROBETTA server for query Rhodopsin.

INTRODUCTION:

Rhodopsin, also called visual purple, pigment-containing sensory protein that converts light into an electrical signal. Rhodopsin is found in a wide range of organisms, from vertebrates to bacteria. In many seeing animals, including humans, it is required for vision in dim light and is located in the retina of the eye—specifically, within the tightly packed disks that make up the outer segment of the retina's photoreceptive rod cells, which are specially adapted for vision under low-light conditions.

The Robetta server provides automated tools for protein structure prediction and analysis. For structure prediction, sequences submitted to the server are parsed into putative domains and structural models are generated using either comparative modeling orde novo structure prediction methods. If a confident match to a protein of known structure is found using BLAST, PSI-BLAST, FFAS03 or 3D-Jury, it is used as a template for comparative modeling. If no match is found, structure predictions are made using the de novo Rosetta fragment insertion method. Experimental nuclear magnetic resonance (NMR) constraints data can also be submitted with a query sequence for RosettaNMR de novo structure determination. Other current capabilities include the prediction of the effects of mutations on protein–protein interactions using computational interface alanine scanning. The Rosetta protein design and protein–protein docking methodologies will soon be available through the server as well.

METHODOLOGY:

1. Open homepage for Robetta (URL: https://robetta.bakerlab.org/)
2. Complete registration
3. Submit FASTA sequence for kinase.
4. Observe and interpret results.

OBSERVATION:

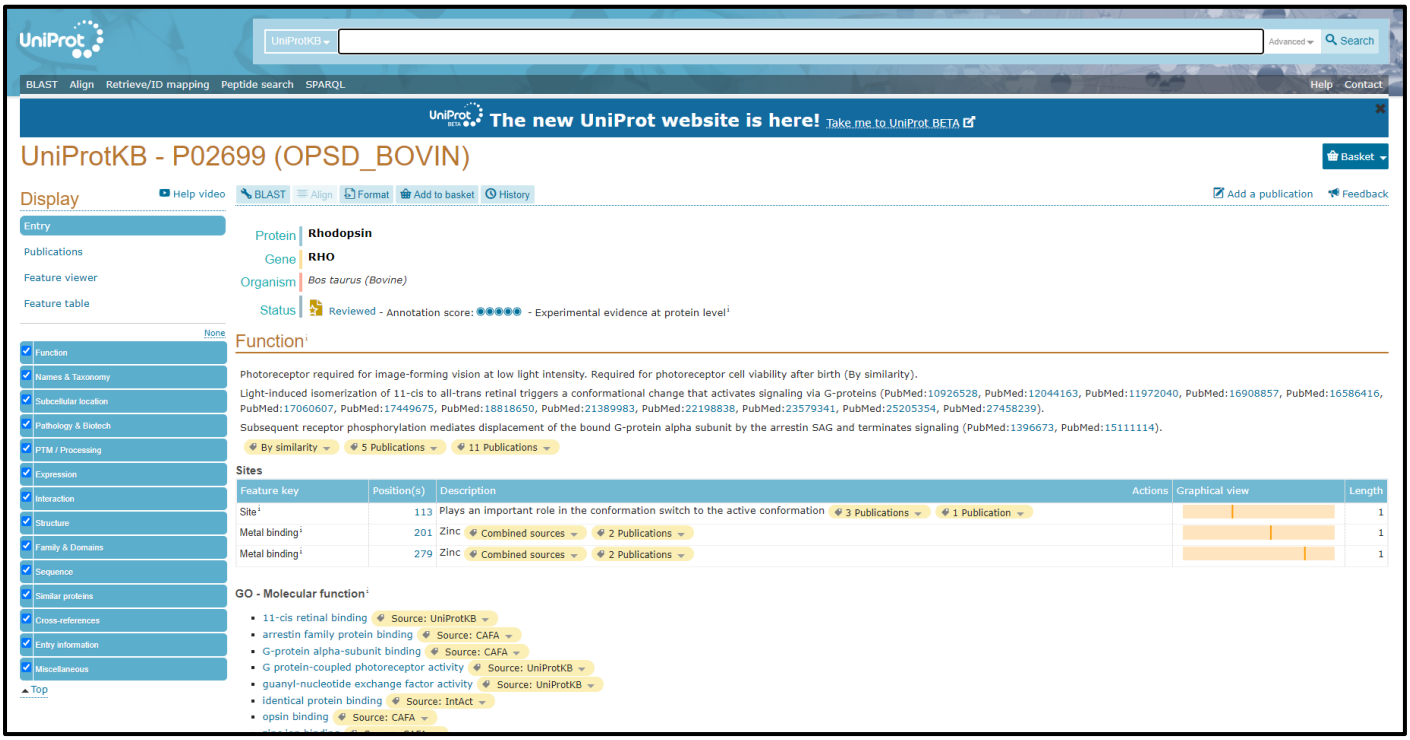


Fig1. Result page for Rhodopsin in UniProt database

```
>sp|P02699|OPSD_BOVIN Rhodopsin OS=Bos taurus OX=9913 GN=RHO PE=1 SV=1
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
EGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAV
YNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA
```

Fig2. FASTA sequence for Rhodopsin

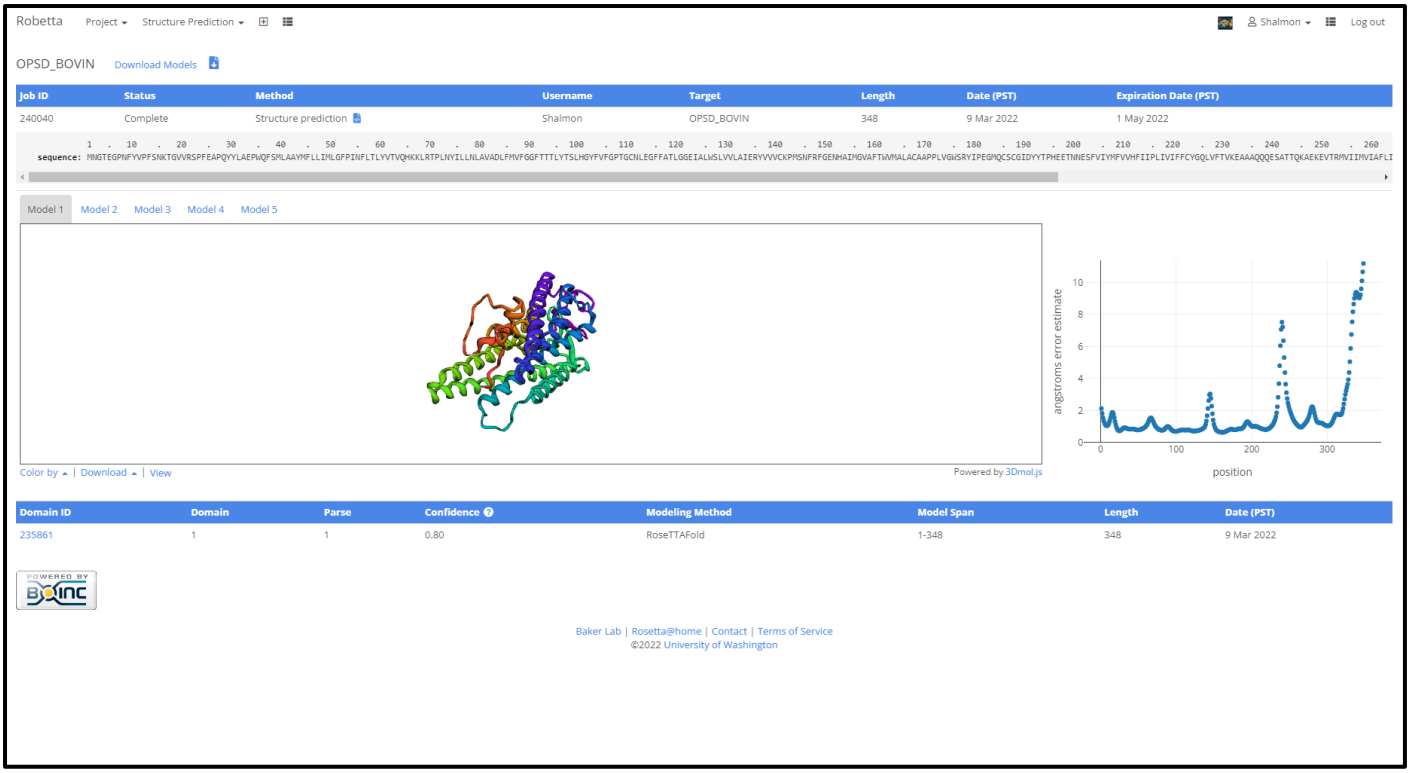


Fig3. Model 1 with atom co-ordinates

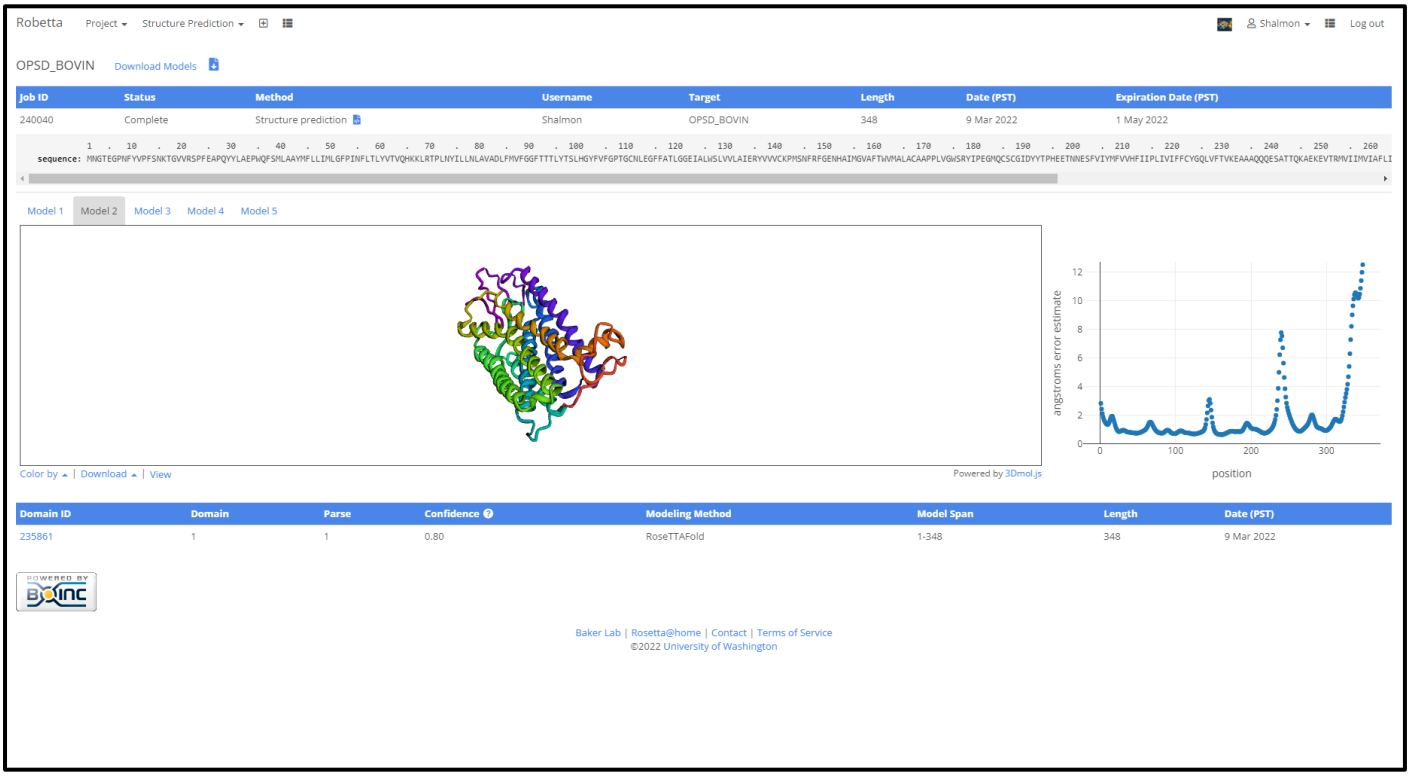


Fig4. Model 2 with atom co-ordinates

Fig5. Model 3 with atom co-ordinates

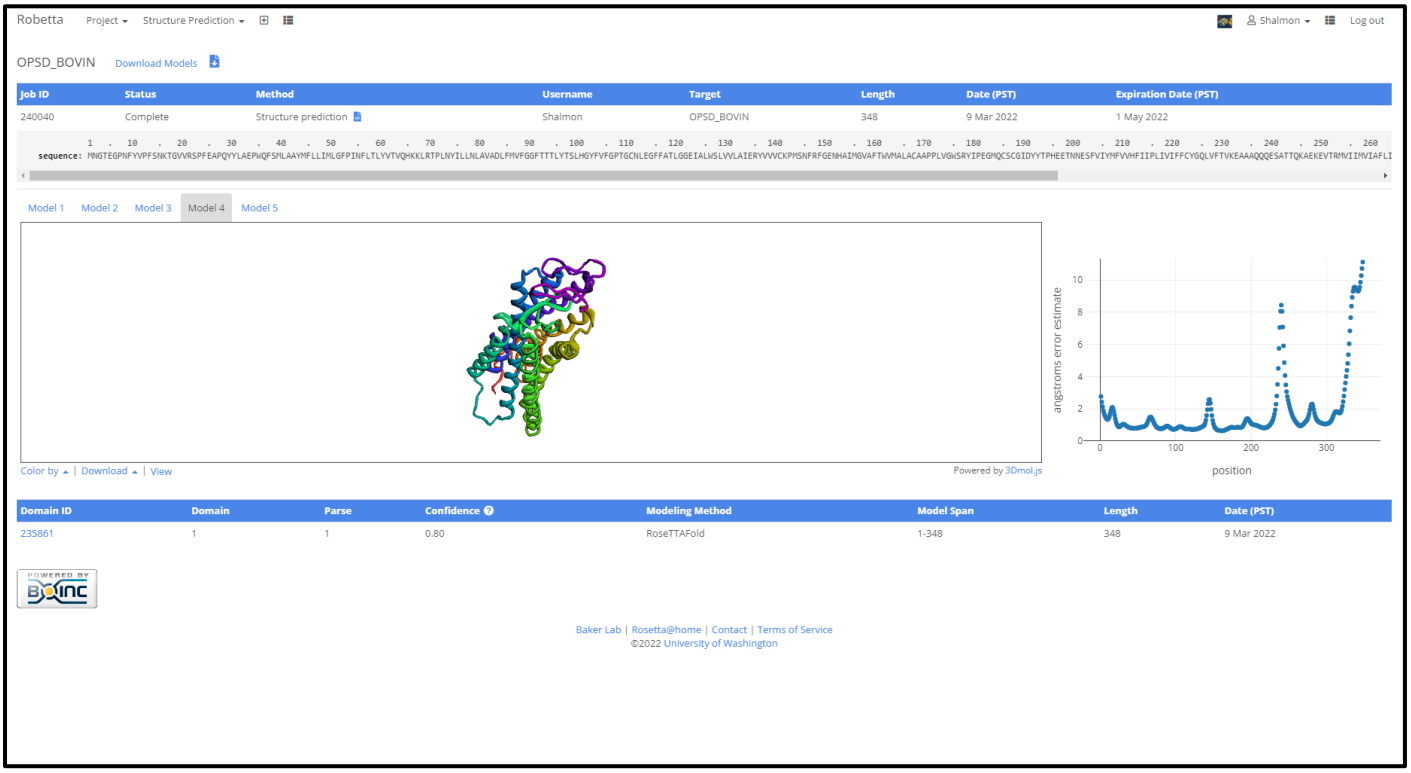


Fig6. Model 4 with atom co-ordinates

Fig7. Model 5 with atom co-ordinates

RESULT:

Robetta was used to predict the tertiary structure of Rhodopsin based on ab-initio approach.

CONCLUSION:

Thus, Robetta can be used to predict tertiary structures of proteins by ab-initio method. These tools give faster results than x-ray or NMR techniques and can be used by researchers to understand protein functions and drug designing.

REFERENCES:

1. Xiong, J. (2008).Tertiary structure prediction. Essential bioinformatics. Cambridge: Cambridge University Press. 214-228.
2. Encyclopædia Britannica, inc. (n.d.). *Rhodopsin*. Encyclopædia Britannica. Retrieved March 18, 2022, from https://www.britannica.com/science/rhodopsin
3. Robetta (2021b). Bakerlab.org. Retrieved March 18, 2022, from https://robetta.bakerlab.org/results.php?id=240040