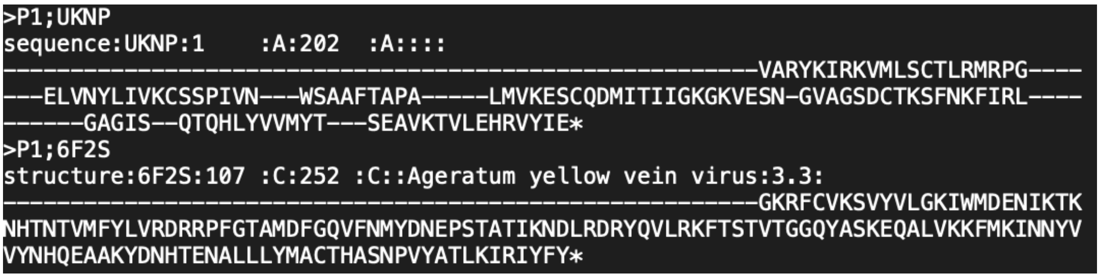
Theodoros Foskolos, Selin Çakmak

Structure Prediction - Structural Bioinformatics 2023

T1026

Question 1:

It is known that similar sequences result in similar structures. To select a suitable template for predicting the structure of T1026, we used the its sequence to query the HHpred server for proteins with high sequence similarity and known structure (Zimmermann et al., 2018). The server generates alignments between target sequence and the sequences of homologous proteins. From this, we select the protein with the highest similarity, 6F2S, as template for homology modelling.



```
>P1;UKNP
sequence:UKNP:1    :A:202  :A::::
---------------------------------------------------VARYKIRKVMLSCTLRMRPG----
---ELVNYLIVKCSSPIVN---WSAAFTAPA-----LMVKESCQDMITIIGKGKVESN-GVAGSDCTKSFNKFIRL----
------GAGIS--QTQHLYVVMYT---SEAVKTVLEHRVYIE*
>P1;6F2S
structure:6F2S:107 :C:252 :C::Ageratum yellow vein virus:3.3:
---------------------------------------------------GKRFCVKSVYVLGKIWMDENIKTK
NHTNTVMFYLVRDRRPFGTAMDFGQVFNMYDNEPSTATIKNDLRDRYQVLRKFTSTVTGGQYASKEQALVKKFMKINNYV
VYNHQEAAKYDNHTENALLLYMACTHASNPVYATLKIRIYFY*
```

Figure 1: Alignment (.pir) file of target sequence (UKNP) and template sequence (6F2S).

We capitalized the name of the PDB file so that the MODELLER script can recognize them and they correspond to the names in the alignment files. For running the build_model.py script we used 6F2S for the "knowns" variable and 'UKNP' for "sequence".

The DOPE (discrete optimized protein energy) score assesses the energy of the homology models, with lower scores indicating higher energetic stability. The following scores obtained from the homology models:

- Model 1 (UKNP.B99990001) | **DOPE**: -10087.57715
- Model 2 (UKNP.B99990002) | **DOPE**: -10082.03223
- Model 3 (UKNP.B99990003) | **DOPE**: - 9786.91406
- Model 4 (UKNP.B99990004) | **DOPE**: - 9859.04102
- Model 5 (UKNP.B99990005) | **DOPE**: - 9638.08105

We find the lowest DOPE score for the first model, which is therefore the most energetically stable structure.

Question 2:

In LGA (Zemla et al., 2003), the GDT_TS (global distance test total score) indicates the similarity between the predicted protein structure and the solution structure. It ranges from 0 to 100, with higher scores indicating a higher similarity between the structures. We find the following scores for our five models resulting from MODELLER and AlphaFold, when compared to the solution structure 6S44:

- Model 1 (UKNP.B99990001) | **GDT_TS**: 71.522 | **DOPE**: -10087.57715
- Model 2 (UKNP.B99990002) | **GDT_TS**: 69.130 | **DOPE**: -10082.03223
- Model 3 (UKNP.B99990003) | **GDT_TS**: 70.870 | **DOPE**: - 9786.91406
- Model 4 (UKNP.B99990004) | **GDT_TS**: 67.609 | **DOPE**: - 9859.04102
- Model 5 (UKNP.B99990005) | **GDT_TS**: 65.870 | **DOPE**: - 9638.08105
- AlphaFold | **GDT_TS**: 93.836

We find the highest structure similarity under our models for the first one with a total score of 71.522. This is in agreement with the GDT_TS results and we therefore find the first model to be the most accurate for predicting the structure of this protein.

In comparison, we find that the AlphaFold model has lower structural difference to the solution than all our homology models with a GDT_TS of 93.836 and is therefore the most accurate. This agrees with our expectations because AlphaFold is an advanced deep learning-based prediction algorithm that takes more than the available experimental data into account (subsequently explained), whereas MODELLER relies heavily on the experimental information and the template structure.

Question 3:

The global distance test columns in the LGA output indicate the number of residues within specified distance thresholds in the structures. Based on this, we observe that the residues 1-3 of T1026 are modelled most poorly.

These correspond to residues 57-59 in both the solution structure and the AlphaFold model. While the AlphaFold model is modelled better in this region, it still shows slight deviations from the solution structure 6S44. These regions are located at the terminus of the proteins, as shown in Figure 2 (left), and the flexibility of this loop region may explain poor modelling results. Flexible loop regions are a common challenge in structural predictions of proteins. It is worth noting that regions in the middle of the protein structure (see Figure 2, right) are accurately modelled.

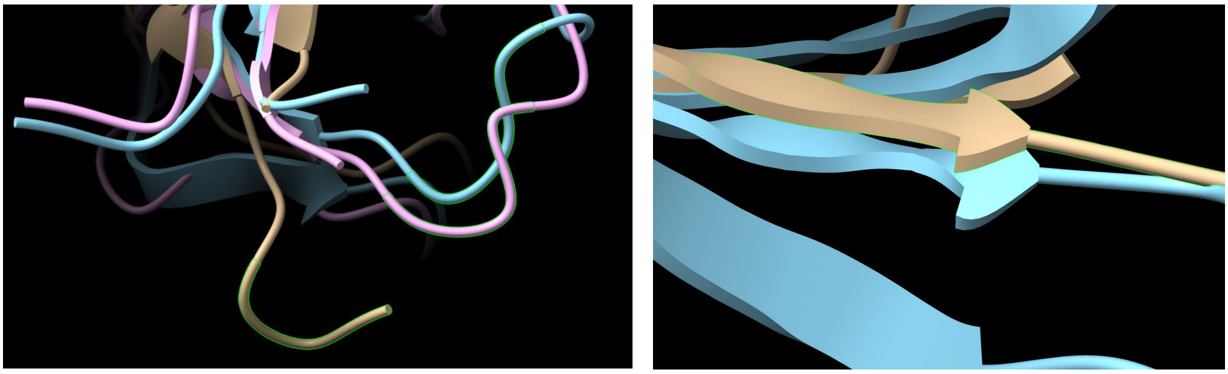


Figure 2: Structures of UKNP.B99990001 (brown), AlphaFold (purple), 6S44 (blue) matched in Chimera X. The selected green regions shows the poorly modelled (left, position 1-3) and accurately modelled (right, position 112-115) residues.

Question 4 [112 words]:

MODELLER uses the sequence alignment between the target and template protein to determine the homology model. Gaps in the template structure are regions where no information is available to guide the homology modelling. One way MODELLER deals with unaligned regions (gaps) is by filling in the missing residues based on the surrounded sequence information, as we can see in Figure 3. where it uses amino acids that fit the β-sheet structure. Another way this algorithm deals with these regions is to treat them as unmodelled loops (Webb et al., 2016). This is an important step since inserting residues into big continuous gap regions could lead to significant inaccuracies in the final model.

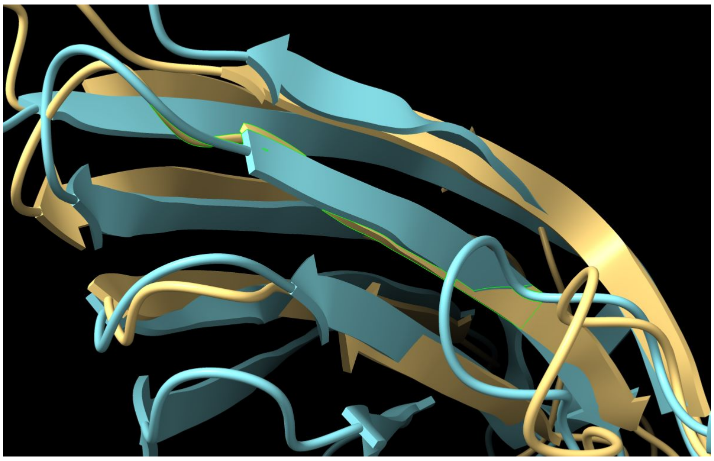


Figure 3: Position of the first gap in sequence alignment (residue 21-27) in the homology model UKNP.B99990001 (brown) and 6S44 (blue).

Question 5 [120 words]:

Flanking terminal regions at the termini of proteins are often flexible and disordered, which makes them difficult to detect experimentally with X-ray crystallography or NMR-spectroscopy. Nevertheless, these regions can play important functional roles such as interactions with other molecules. This may explain why our solution 6S44 did not resolve the specific termini region. However, since AlphaFold is trained on a variety of structural data including PDB, it is able to predict these regions. This is depicted in the matched structures in Figure 4. In CASP, AlphaFold flanking regions are excluded to prevent local biases, focus on the core regions of the protein where most of its structure is formed and increase accuracy in the overall assessment of the protein structure.

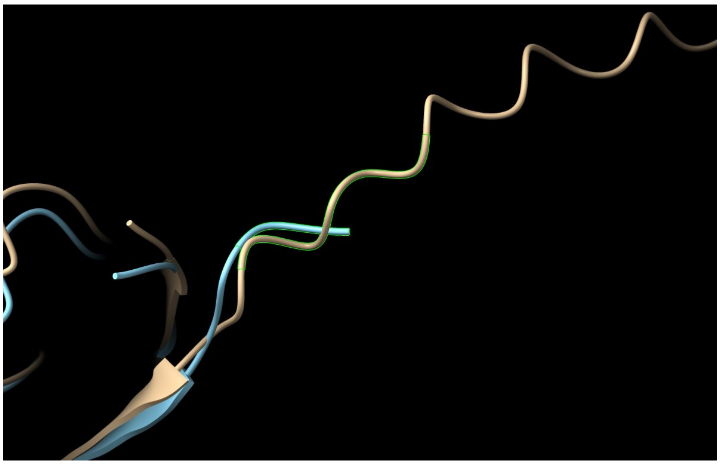
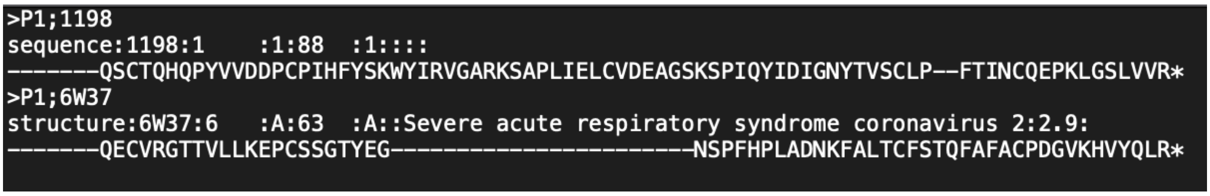


Figure 4: Flanking regions in matched structures of AlphaFold model (brown) and 6S44 solution structure (blue).

T1064

Question 6:

To select a suitable template for predicting the structure of T1064, we used the same approach as for T1026. First we query the target sequence in the HHpred server for proteins with high sequence similarity and known structure. The protein with the highest similarity, 6W37, is used as template for homology modelling.



```
>P1;1198
sequence:1198:1    :1:88  :1::::
-------QSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIELCVDEAGSKSPIQYIDIGNYTVSCLP--FTINCQEPKLGSLVVR*
>P1;6W37
structure:6W37:6   :A:63  :A::Severe acute respiratory syndrome coronavirus 2:2.9:
-------QECVRGTTVLLKEPCSSGTYEG-----------------------NSPFHPLADNKFALTCFSTQFAFACPDGVKHVYQLR*
```

Figure 5: Alignment (.pir) file of target sequence and template sequence (6W37).

We capitalized the PDB file name so that they correspond to the names in the alignment files. For running the build_model.py script we used 6W37 for the "knowns" variable and '1198' for "sequence".

- Model 1 (1198.B99990001) | **DOPE**: -5219.10547
- Model 2 (1198.B99990002) | **DOPE**: -5464.77930
- Model 3 (1198.B99990003) | **DOPE**: -5474.73145
- Model 4 (1198.B99990004) | **DOPE**: -5081.02002
- Model 5 (1198.B99990005) | **DOPE**: -5622.18066

We find the lowest DOPE score for the fifth model, which is therefore the most suitable (energetically favorable) model.

Question 7:

Both modelling approaches (AlphaFold and MODELLER) perform worse with this target sequence. Firstly, the higher complexity of the structure can present a challenge for prediction for both. The native protein contains two chains, but only A chain is being modelled here. This means that interactions between the chains might not be taken into account, which can affect the accuracy of predictions.

Additionally, the two target sequences have different alignment scores with their corresponding templates. The first target T1026 has a higher alignment score with the template (30.04) than the second target T1064 (22.82). This suggests that T1026 has higher sequence similarity to the template than T1064.

Finally, since the target T1064 is a (relatively) recently studied molecule, there might be limited experimental evidence on its structure. The more complex structure and the limited data make finding an appropriate template structure difficult since the available data might not be sufficient to capture the features of the target. In contrast to MODELLER, AlphaFold uses physical, geometric and evolutionary information to make structure predictions (Jumper et al., 2021). Here, a drop in accuracy can result from not having a sufficient amount of homologous sequences in the initial multiple sequence alignment (MSA depth).

Question 8:

The GDT_TS scores of 5 AlphaFold models compared to the solution structure:

- AlphaFold 1 | **GDT_TS**: 80.392
- AlphaFold 2 | **GDT_TS**: 16.176
- AlphaFold 3 | **GDT_TS**: 17.647
- AlphaFold 4 | **GDT_TS**: 14.216
- AlphaFold 5 | **GDT_TS**: 80.392

We observed 2 missing residues (see Figure 6) in the solution structure which affected the LGA alignment and can decrease the similarity score. Therefore, we introduce gaps at the corresponding positions on the AlphaFold models to achieve a higher similarity score of the structures and a higher GDT_TS score. The difference between the accuracy of the structures derived from Modeller and AlphaFold is lower than for the first target, T1026. The highest score within the models is 80.392 and is found for both the first and fifth model. These AlphaFold models seem less stable than the ones for the previous target with lower GDT_TS scores and high variation between the five model results, ranging from 14 to 80.

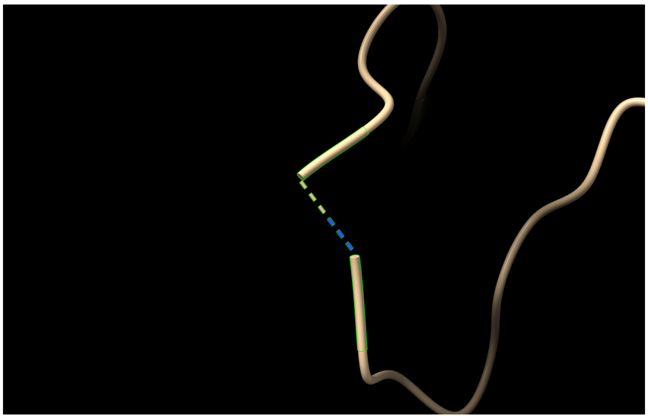


Figure 6: Missing (unmodelled) residues 65 and 66 in the solution structure 7JTL.

In real-world applications, missing solution structures make it hard to validate the performance of AlphaFold models, as we cannot calculate structural similarity of the structures to the native states of the target protein. Even if the AlphaFold model performs well for some proteins it cannot be used an indicator for the rest.

Question 9:

AlphaFold is a deep learning-based structure prediction algorithm that handles limited data by not only using PDB information, but also incorporating physical and geometric inductive bias to learn missing physical contexts, such as hydrogen bond formation. In contrast, MODELLER only takes the experimental data and the template structure into account.

The observed residue distances (LGA) between AlphaFold models are lower than the MODELLER ones, indicating higher similarity for both targets. The distances were also lower for all type of models generated for the easy target in comparison with the hard one. This is also evident from the RMSD scores after matching the models together in Chimera. However, AlphaFold experiences drops in accuracy when dealing with complex or newly studied proteins, since it relies heavily on the initial multiple sequence alignment. If not enough homologous sequences are available in the MSA, its performance decreases.

Combining these information and also the issues stated in the previous questions, we conclude that MODELLER is outperformed by AlphaFold in accuracy and handling limited data, due to its feature to incorporate additional information when predicting a structure.

Contributions

Contributor Role	
Formal Analysis	Theodoros Foskolos, Selin Çakmak
Methodology	Theodoros Foskolos, Selin Çakmak
Project Administration	Theodoros Foskolos, Selin Çakmak
Software	Theodoros Foskolos, Selin Çakmak
Visualization	Theodoros Foskolos, Selin Çakmak
Writing	Theodoros Foskolos, Selin Çakmak

References

A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. J Mol Biol. 2018 Jul 20. https://doi.org/10.1016/j.jmb.2017.12.007

Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003 Jul 1;31(13):3370-4. https://doi.org/10.1093/nar/gkg571 .

Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. Current protocols in bioinformatics, 54(1), 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3