# Grundlagen der Bioinformatik

SoSe 2022

Tutor: Theresa/ Mathias

| 1 | 2 | 3 | $\sum$ |
|---|---|---|---|
| | | | |

Marina Dittschar & Clarissa Auckenthaler

## Blatt 12

(Abgabe am 28.07.2022)

# Theoretical Assignments

### Task 1: Fractional Overlap of Segments (SOV)(8)

Given:

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $S_{obs}$ | C | C | C | E | E | E | E | E | E | C | C | C | E | E | E | E | E | E | C | C |
| $S1_{pred}$ | C | C | C | C | E | E | E | C | C | C | C | C | C | E | E | E | E | C | C | C |
| $S2_{pred}$ | C | C | C | E | C | E | E | E | C | C | C | E | E | E | C | E | C | E | C | C |

Calculation of SOV for $S_{obs}$ and $S1_{pred}$:

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $S_{obs}$ | C | C | C | E | E | E | E | E | E | C | C | C | E | E | E | E | E | E | C | C |
| $S1_{pred}$ | C | C | C | C | E | E | E | C | C | C | C | C | C | E | E | E | E | C | C | C |

$S(C) = \{(CCC, CCCC), (CCC, CCCCCC), (CC, CCC)\}$

$S'(C) = \emptyset$

$S(E) = \{(EEEEEE, EEE), (EEEEEE, EEEE)\}$

$S'(E) = \emptyset$

$N(C) = 3 + 3 + 2 = 8$

$N(E) = 6 + 6 = 12$

$SOV(C) = 100 * \frac{1}{8}[\frac{3+1}{4} * 3 + \frac{3+1}{6} * 3 + \frac{2+1}{3} * 2] = 0.875$

$SOV(E) = 100 * \frac{1}{12}[\frac{3+1}{6} * 6 + \frac{4+2}{6} * 6] = 83.33$

$SOV = 100 * \frac{1}{8+12}[\frac{3+1}{4} * 3 + \frac{3+1}{6} * 3 + \frac{2+1}{3} * 2 + \frac{3+1}{6} * 6 + \frac{4+2}{6} * 6] = \frac{100}{20}[3+2+2+4+6] = 85$

$Q_3 = \frac{15}{20} * 100 = 75$

Calculation of SOV for $S_{obs}$ and $S2_{pred}$:

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $S_{obs}$ | C | C | C | E | E | E | E | E | E | C | C | C | E | E | E | E | E | E | C | C |
| $S2_{pred}$ | C | C | C | E | C | E | E | E | C | C | C | E | E | E | C | E | C | E | C | C |

$S(C) = \{(CCC, CCC), (CCC, CCC), (CC, CC)\}$

$S'(C) = \emptyset$

$S(E) = \{(EEEEEE, E), (EEEEEE, EEE), (EEEEEE, EEE), (EEEEEE, E), (EEEEEE, E)\}$

$S'(E) = \emptyset$
$N(C) = 3 + 3 + 2 = 8$
$N(E) = 6 + 6 + 6 + 6 + 6 = 30$
$SOV(C) = 100 * \frac{1}{8}[\frac{3+0}{3} * 3 + \frac{2+1}{4} * 3 + \frac{2+0}{2} * 2] = 87,5$
$SOV(E) = 100 * \frac{1}{30}[\frac{1+0}{6} * 6 + \frac{3+1}{6} * 6 + \frac{2+1}{7} * 6 + \frac{1+0}{6} * 6 + \frac{1+0}{6} * 6] = 31.9$
$SOV = 100 * \frac{1}{8+30}[\frac{3+0}{3} * 3 + \frac{2+1}{4} * 3 + \frac{2+0}{2} * 2 + \frac{1+0}{6} * 6 + \frac{3+1}{6} * 6 + \frac{2+1}{7} * 6 + \frac{1+0}{6} * 6 + \frac{1+0}{6} * 6] =$
$\frac{100}{38}[3 + 2.25 + 2 + 1 + 4 + 2.57 + 1 + 1] = 44.26$
$Q_3 = \frac{15}{20} * 100 = 75$

Even if the $Q_3$ for both has the same value of 75% the SOV for the two predictions is different. For the $S1_{pred}$ the SOV is 85% and for $S2_{pred}$ the SOV is 44.26%. This means that $S1_{pred}$ is closer to the $S_{obs}$ than $S2_{pred}$.

## Practical Assignments

### Task 2: Predict secondary structure of proteins (8)

We used Python 3.8.8, and the following libraries **sys, getopt, and numpy**.

We used the tools "PredictProtein" [1] and "Proteus2" [3] to obtain the secondary structure predictions of the given protein. We obtained the sequence predictions from the "featureString" field for "PROFsec" of the XML-exported output of "PredictProtein" and from the "Predicted Complete Secondary Structure" section from "Proteus2". We saved the information in two files named "auckenthaler_dittschar_PHD_featurestring.txt" and "auckenthaler_dittschar_proteus2_prediction.txt".
Enter the following code in the command line to run the file:

```
python auckenthaler_dittschar_q3_comparison.py -a auckenthaler_dittschar_PHD
_featurestring.txt -b auckenthaler_dittschar_proteus2_prediction.txt
-r trueSecStructure.txt
```

For the Q3 values, we get a score of 0.7516 for "PredictProtein" and a score of 0.8824 for "Proteus2". Therefore, we can say that for "Proteus2", the proportion of exactly matching structure predictions is higher and the prediction therefore is more accurate.

### Task 3: 3D protein structure prediction (4)

We used the tools "Phyre2" [3], "I-Tasser" [7] and "AlphaFold" [2]. We shared the application of these tools between three groups to reduce runtime, Lea Heinen and Marit Bockstedte applied "I-Tasser" and Nadja Buttke "AlphaFold", while we applied "Phyre2".

You can see the results of the algorithms in Figures 1, 2 and 3 According to both "Phyre2" and "I-Tasser", the structure is the ATP-dependent DNA helicase DinG (see both results: `http://www.sbg.bio.ic.ac.uk/phyre2/phyre2_output/cf24946fc4e420ec/summary.html`

and `https://seq2fun.dcmb.med.umich.edu//I-TASSER/output/S696591/`). Helicase is an enzyme that unwinds DNA strands [6]. Visually comparing the output results shows that all three results are very similar, both in global structure and proportion of alpha helices for example. We can see in the "Phyre2"-report, that the reported structure has a high confidence (100.0) and %i.d. score (99). Following the "I-Tasser" result, the protein forms a "complex with ssDNA" in Escheria coli [5], while following "Phyre2", it forms a "complex with ssDNA and ADPBeF" [4]. The results are therefore very similar and even though there is no predicted protein suggested in the resuls of "AlphaFold", structurally it also comes to the same conclusions, except for a part at the end of the protein sequence where there is a part of the protein that is not in an alpha helix or a beta sheet.

(a) Predicted 3D structure for the amino acid sequence by:
Phyre2: c6fwsB

(b) Predicted 3D structure for the amino acid sequence by:
I-Tasser: 6fwrA

(c) Predicted 3D structure for the amino acid sequence by:
AlphaFold2

Figure 1: Results of running the three programs.

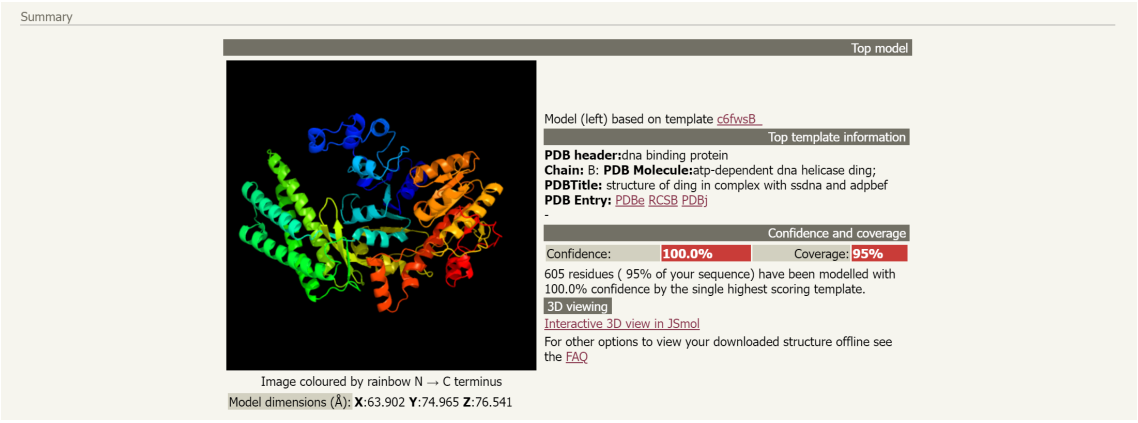


Figure 2: Screenshots result sheet Phyre2


Figure 3: Screenshots result sheet I-Tasser


Figure 4: RCSB entry for the structure 6FWS

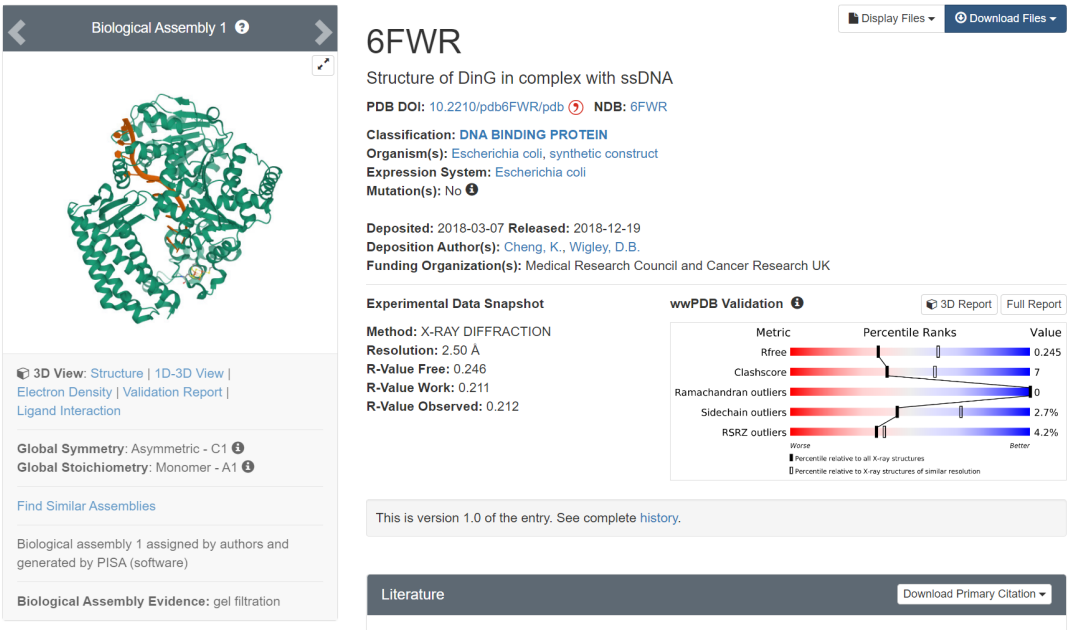


Figure 5: RCSB entry for the structure 6FWR

## References


[1] Michael Bernhofer, Christian Dallago, Tim Karl, Venkata Satagopam, Michael Heinzinger, Maria Littmann, Tobias Olenyi, Jiajun Qiu, Konstantin Schütze, Guy Yachdav, Haim Ashkenazy, Nir Ben-Tal, Yana Bromberg, Tatyana Goldberg, Laszlo Kajan, Sean O'Donoghue, Chris Sander, Andrea Schafferhans, Avner Schlessinger, Gerrit Vriend, Milot Mirdita, Piotr Gawron, Wei Gu, Yohan Jarosz, Christophe Trefois, Martin Steinegger, Reinhard Schneider, and Burkhard Rost. PredictProtein - Predicting Protein Structure and Function for 29 Years. Nucleic Acids Research, 49(W1):W535–W540, 05 2021.

[2] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. Nature Methods, pages 1–4, 2022.

[3] Scott Montgomerie, Joseph A Cruz, Savita Shrivastava, David Arndt, Mark Berjanskii, and David S Wishart. Proteus2: a web server for comprehensive protein structure prediction and structure-based annotation. Nucleic acids research, 36(suppl_2):W202–W209, 2008.

[4] RCSB. Rcsb pdb - 6fwr: Structure of ding in complex with ssdna. `https://www.rcsb.org/structure/6fwr`, 2022.

[5] RCSB. Rcsb pdb - 6fws: Structure of ding in complex with ssdna and adpbef. `https://www.rcsb.org/structure/6fws`, 2022.

[6] Y Wu. Unwinding and rewinding: double faces of helicase? j nucleic acids 2012: 140601, 2012.

[7] Wei Zheng, Chengxin Zhang, Yang Li, Robin Pearce, Eric W Bell, and Yang Zhang. Folding non-homologous proteins by coupling deep-learning contact maps with i-tasser assembly simulations. Cell reports methods, 1(3):100014, 2021.