Structural integrity of early life protein analogues

Shamim Ekramullah — Wbn215

January 24, 2025

Preface

During my bachelors degree in biochemistry, I had the honour of working with Associate Professor Tue Hassenkam from the GLOBE Institute at KU on two seperate occasions. Both of these projects served as an introduction to the origin of life debate. [2, 3]

I saw this project as an opportunity to do some genuinely interesting work and research. Thus, I wanted it to serve as an extension as some of my previous work, by researching a protein analogue from the origin of life debate.

Introduction

The origin of life, is highly debated topic among scientists from all fields. As there is little substance in the form of evidence of the first organisms, the debate is highly theoretical and skeptical in nature. This has led to a division of the scientific community into two camps. This is very much a debate of what came first, the chicken or the egg - or more scientifically - metabolism or Darwinism? I have covered this debate thoroughly in previous work, but to summarise:

One camp argues that deep sea hydrothermal vent systems in a sightly acidic (pH 6) hadean ocean, would provide the best conditions for the emergence of prebiotic metabolic analogues. This is due to the nature of the systems being rich in mineral deposits and reaction interfaces. Another popular anectode is the precense of gradients, both temperature and pH gradients, which could be used to drive chemical reactions - much akin to that of the electron transport chain in modern mitochondria. The proposed idea is that, because minerals and metal-ions play a crucial role in modern lifeforms across all levels of complexity, it makes perfect sense for life to have emerged where such compounds would be found in abundance.

On the other side of the debate are those who believe that life emerged on the earths surface during the hadean era, in little pools around hydrothermal fields. It is argued that poly-peptides are likely to be the first remnants of what would later become genetic material. Hydrothermal fields offer the solvent to undergo wet-dry cycles, which have been shown to foster condensation reactions, which is absolutely crucial in driving polymerisation. [2]

I saw a couple of different opportunities to cater a structural bioinformatics project to accomodate this debate. Firstly I was drawn to the prospect of researching ancient proteins related to the hydrothermal field theory. Here ribozymes would have been an ideal choice, but I actually decided on a different approach.

In my second project with Associate Professor Tue Hassenkam, I worked on a review of the deep sea hydrothermal vent theory. In this review, there is a section which discusses the idea of ferredoxins emergence in said environment.[3] Ferredoxins serve as electron carriers, and the idea is that they were present to drive prebiotic metabolism using gradients. Therefore, I sought to research the "structural integrity" of a modern ferredoxin. Naturally, deep sea hydrothermal vents are situated in very extreme and potentially hostile environments. Thus, ferredoxins during this time-period would have to be stable under a lot of varying conditions, hereamongst pressure, temperature variation and pH variation.[3]

In the coming sections I will be presenting my protein of choice, the "experiments" I conducted and the somewhat surprising results I found.

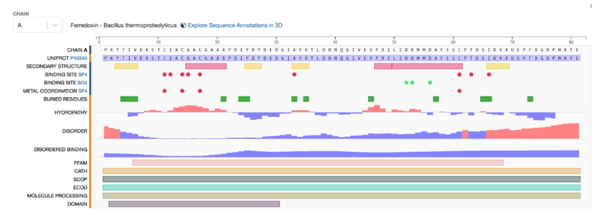
CHAIN
A
Ferredoxin - Bacillus thermoproteolyticus
Explore Sequence Annotations in 3D
0
10
20
30
40
50
60
70
80
CHAIN A
P K Y T I V D K E T C I A C G A C G A A A P D I Y D Y D E D G I A Y V T L D D N Q G I V E V P D I L I D D M M D A F E G C P T D S I K V A D E P F D G D P N K F E
UNIPROT P10245
SECONDARY STRUCTURE
BINDING SITE SF4
BINDING SITE SO4
METAL COORDINATION SF4
BURIED RESIDUES
HYDROPATHY
DISORDER
DISORDERED BINDING
PFAM
CATH
SCOP
ECOD
MOLECULE PROCESSING
DOMAIN

Figure 1: Primary and secondary structure overview of 1IQZ

1 Selecting a structure from the PDB

1.1 How to create Sections and Subsections

As previously mentioned I first considered working with ribozymes. However, I found it a greater challenge than expected to find a ribozyme that fulfilled all of the given criteria.

We wanted a protein in a crystal structure of fairly small protein in a good resolution (¡2Å), which absolutely was not the infamous lysozyme.

Eventually, I found a ferredoxin from Bacillus thermoproteolyticus which was perfect for this project. I found this by searching for molecules in the protein data bank (PDB), which had ferredoxin in its Uniprot name.

Keiichi Fukuyama et al.[4] Published two seperate articles back in 2002, on two forms of *Bacillus thermoproteolyticus* ferredoxin (BtFd) in 0.92Å and 1Å respectively for form I (PDBID: 1IQZ) and form II (PDBID: 1IR0). I considered anything below a resolution of 2Å a viable resolution, so both of these forms would be viable options. Ultimately, I chose form 1, which consists of 81 residues.

From the secondary structure prediction on rcsb.org, I was able to get a quick overview of the secondary structure, which consists of two alpha helices and four beta sheets, as seen by Figure 1.

A preview of the tertiary structure can be found in the rcsb 3D structure preview. Here it is very apparent to see the Fe-S ligand that sits in the active site of the protein (Figure 2 + 3).

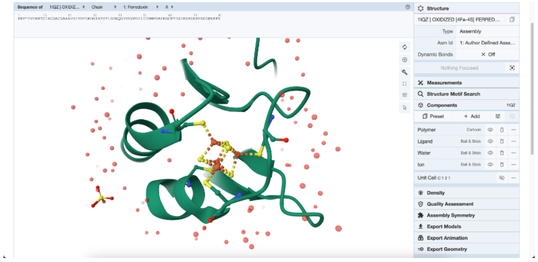

Figure 2: Tertiary structure from PDB

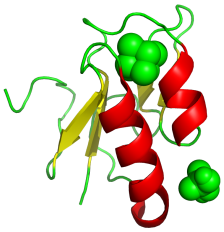

Figure 3: Tertiary structure overview in pymol

I used Biopython to retrieve the .pdb file of 1IQZ as seen from the download_pdb_file() in the script protein.py.

2 Primary and secondary structure analysis

Following the retrieval of the 1IQZ structure, I conducted an analysis of its primary and secondary structural features using Biopython's PPBuilder module. This allowed for the identification of any chain breaks and verification of secondary structure elements such as helices and beta sheets.

2.1 Primary Structure Analysis

The primary structure analysis revealed no chain breaks in the sequence of residues within the PDB file. This result confirms that the structure is fully mapped, ensuring its suitability for further analyses.

Chain breaks in protein structures refer to discontinuities in the residue sequence as represented in a PDB file. These breaks occur when residue numbering is non-sequential or when key backbone atoms, such as the alpha carbon (CA), are missing. Ensuring the absence of chain breaks is critical to maintain the structural and functional integrity of the protein model.

2.2 Secondary Structure Analysis

The secondary structure analysis confirmed the findings from the initial prediction on RCSB.org. The structure consists of two alpha helices and four beta sheets, with the following residue ranges:

```
Secondary Structure:
Alpha-Helices:
Helix from ('A', (' ', 16, ' ')) to ('A', (' ', 21, ' '))
Helix from ('A', (' ', 48, ' ')) to ('A', (' ', 61, ' '))

Beta-Sheets:
Sheet from ('A', (' ', 3, ' ')) to ('A', (' ', 7, ' '))
Sheet from ('A', (' ', 25, ' ')) to ('A', (' ', 28, ' '))
Sheet from ('A', (' ', 33, ' ')) to ('A', (' ', 36, ' '))
Sheet from ('A', (' ', 66, ' ')) to ('A', (' ', 70, ' '))
```

The designation 'A' in the output refers to the sole chain of the protein, which is negligible for this analysis.

This detailed understanding of the primary and secondary structure ensures the validity of the 1IQZ model and provides a solid foundation for subsequent studies, including mutational analysis.

3 Residue-Ligand analysis

This phase of the project marked a critical point where I implemented strategies to evaluate the structural integrity of BtFd (1IQZ). By analyzing the tertiary structure visually in PyMOL, it became evident that several secondary structure elements resided within the active site. Consequently, I focused on disrupting the interaction between the ligand (SF4) and BtFd as a key avenue for investigation.

3.1 Algorithmic Approach to Residue Selection

Rather than directly altering secondary structures, I developed an algorithm to automatically identify residues in proximity to SF4. The function `get_interacting_residues()` achieves this by applying two criteria:

- **Interaction Distance Threshold**: Residues within a 5Å radius of the ligand.
- **Bond Angle Cutoff**: Residues forming a bond angle of 120 degrees or less relative to the ligand.

To calculate bond angles, I utilized a helper function `calculate_angle`, which determines the angle between two vectors in degrees. This automated approach provided a comprehensive list of all residues interacting with SF4.

3.2 Relevance of Residue-Ligand Interactions

Residue-ligand interactions are fundamental to protein functionality, particularly in structural bioinformatics. These interactions are influenced by both spatial proximity and geometric orientation, with the latter playing a crucial role in defining interaction strength and specificity[8]. Incorporating an angle threshold into residue selection enhances the precision of identifying biologically and chemically significant interactions.[6]

The use of angle thresholds in residue selection represents a methodological advancement, ensuring the accurate modeling of protein-ligand interactions. This approach is especially pertinent to origin-of-life research, offering insights into primitive protein functions and bridging the gap between early biochemical systems and contemporary molecular biology.[7, 1]

The list is really quite long. This is a snippet:

- **CYS 61:** Distance 2.27 Å, Angle 108.16°
- **CYS 14:** Distance 2.27 Å, Angle 76.15°
- **CYS 17:** Distance 2.27 Å, Angle 92.27°
- **CYS 11:** Distance 2.30 Å, Angle 101.98°
- **CYS 17:** Distance 3.23 Å, Angle 85.99°
- **CYS 11:** Distance 3.35 Å, Angle 94.01°
- **CYS 61:** Distance 3.37 Å, Angle 101.61°
- **THR 63:** Distance 3.44 Å, Angle 105.47°
- **CYS 14:** Distance 3.44 Å, Angle 73.52°
- **CYS 14:** Distance 3.46 Å, Angle 116.15°

But it is apparent that quite a few of these residues reside in the intervals of the helices and beta sheets, for instance Cys 17 and Cys 61, which reside in the first and second helix respectively.

As I am writing this, I realise that quite a lot of the residues appear multiple times with different values. I assume this is because SF4 consists of multiple atoms, and thus there is an entry for every interaction. This was not intended, but should not present an issue for the following mutations.

The residues selected for mutation could have been selected by hand, but I again wanted to automate the process, so I designed a function which based on distance- and angle threshold and a new variable target_residues, where I was able to specify which residues to prioritise for mutation. The selected residues were stored in a list, whereafter I employed my mutate_sequence_full_coverage to mutate the selected residues to specified residues. In this case I chose to mutate the residues to Ala and Gly as these are the simplest residues and also the oldest. This is relevant as ancient ferredoxin analogues most likely would not have utilised more complex amino acid residues.

The mutation process prints out all mutated residues and saves the mutated sequence as a fasta file.

Selected residues for mutation (snippet):

- CYS 61: 2.27 Å
- CYS 14: 2.27 Å
- CYS 17: 2.27 Å
- CYS 11: 2.30 Å
- CYS 17: 3.23 Å
- CYS 11: 3.35 Å
- CYS 61: 3.37 Å
- THR 63: 3.44 Å
- CYS 14: 3.44 Å
- CYS 14: 3.46 Å
- ILE 12: 3.49 Å
- CYS 14: 3.55 Å

Now, to retain the requested structure of this report, I will spoil that I did a bunch of different mutation-runs, as this first set yielded no significant results.

As the automated algorithmic approach did not significantly alter the active site, or disturb the tertiary structure in a meaningful way, I decided to target the second helix directly, as this is the largest of the two. Thus, I developed a function, mutate_and_save_sequence(), to seperately target specific residues and save a seperate fasta file. I did this twice, once to mutate all residues in the helix to Ala, and a second time to mutate them to Pro, simply out of curiosity.

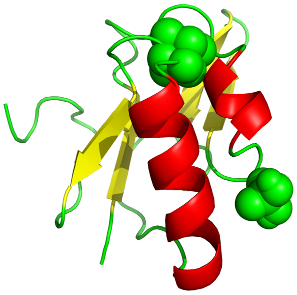

Figure 4: Tertiary structure of algo_mut. As seen, there are no noticeable changes compared to the wild_type structure

4 Tertiary structure prediction

Using ColabFold [5], I generated tertiary structure predictions for the FASTA sequences. This process yielded intriguing yet somewhat unexpected results.

Upon initial inspection, the automated mutant exhibited minimal structural alterations, which was surprising given the mutation's localization to the active site. A rendition of the structure in pymol can be seen on figure 4.

Quantitative Assessment of Structural Alterations To rigorously quantify the structural changes, a custom function was employed to measure residue-ligand distances pre- and post-mutation. Below is a summary of some of the shifts:

Nb. full list can be obtained by running compare_residue_distances().

- Residue ID 11:
  - Wild-Type: CYS - Distance: 2.30 Å
  - Mutant: ALA - Distance: 2.65 Å
- Residue ID 66:
  - Wild-Type: ILE - Distance: 3.70 Å
  - Mutant: ILE - Distance: 2.72 Å
- Residue ID 17:
  - Wild-Type: CYS - Distance: 2.27 Å
  - Mutant: GLY - Distance: 3.01 Å
- Residue ID 61:
  - Wild-Type: CYS - Distance: 2.27 Å
  - Mutant: GLY - Distance: 3.72 Å
- Residue ID 65:
  - Wild-Type: SER - Distance: 4.01 Å
  - Mutant: GLY - Distance: 3.60 Å

These results demonstrate that minor shifts occurred across all residues. "None" denotes interaction distances exceeding the threshold of 5 Å.

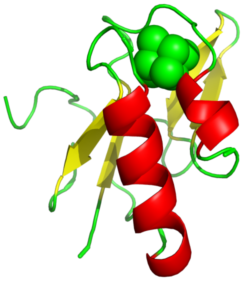

Figure 5: Tertiary structure of helix_mut_ala. No noticeable conformational changes compared to algo_mut and wild type

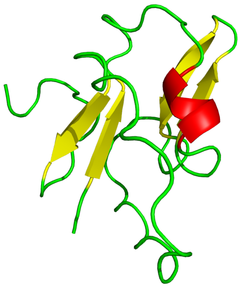

Figure 6: Tertiary structure of helix_mut_pro

Mutation Analysis The helix_ala mutation produced surprising results, as the helical structure persisted despite the mutation. Alanine, being non-polar and relatively neutral, is expected to reduce stabilizing interactions. This unexpected outcome raised concerns about the accuracy of the ColabFold simulation. See figure 5.

To address these concerns, I mutated all residues in the helix to proline as a proof of concept. Proline's unique cyclic structure, combined with its hydrophobic nature, renders it more electrostatically active than alanine.

As anticipated, the conformation of the protein was significantly disrupted, with the helix unfolding almost entirely. See figure 6

5 Evaluation of Predicted Protein Structure Quality

The accuracy and reliability of the predicted protein structures were assessed using two core metrics: the Predicted Local Distance Difference Test (pLDDT) and the Predicted Aligned Error (PAE). These metrics collectively provide a detailed understanding of residue-level accuracy and the global structural integrity of the models.

Detailed Analysis of Confidence Metrics

Algo_mut:

- **pLDDT**: Scores predominantly range above 90, indicating highly reliable local structure predictions (Figure 7).
- **pTM**: High global confidence is reflected in scores exceeding 0.85, suggesting excellent global topology.

helix_mut_ala:

- **pLDDT**: Consistently high scores between 97.04 and 98.05 confirm exceptional accuracy in residue-level predictions (Figure 8).

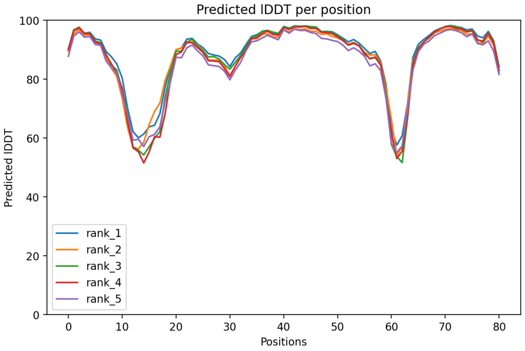

Figure 7: pLDDT overview of algo_mut

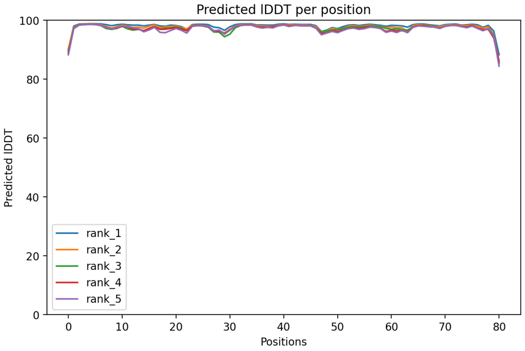

Figure 8: pLDDT overview of helix_mut_ala

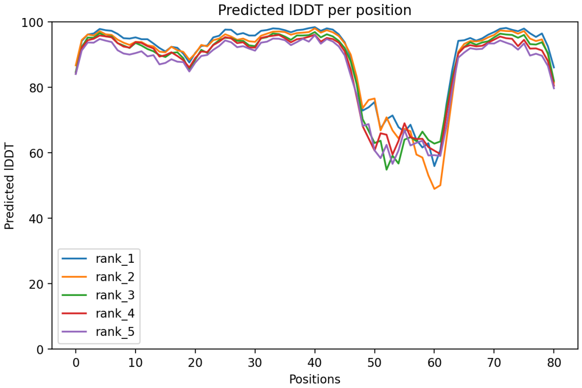

Figure 9: pLDDT overview of helix_mut_pro

- **pTM**: Scores exceeding 0.88 indicate robust confidence in the overall fold and domain orientation.

helix_mut_pro:

- **pLDDT**: Scores range primarily between 90 and 93, demonstrating strong confidence in the local structural features (Figure 9).
- **pTM**: Scores ranging from 0.80 to 0.85 reflect moderate to high confidence in global topology.

Visualization of Confidence Metrics

- **pLDDT Plots**: These plots map the per-residue confidence across the protein sequence. The helix_mutation model exhibits the most residues in the highly confident range (¿90), underscoring its superior local accuracy (Figure 8).
- **PAE Plots**: Heatmaps illustrating inter-residue spatial relationships show minimal deviations for helix_mut_ala, indicating well-defined inter-domain interactions. Conversely, algo_mut and helix_mut_pro display moderate error regions, highlighting uncertainties in domain positioning (Figures 10 and 11).

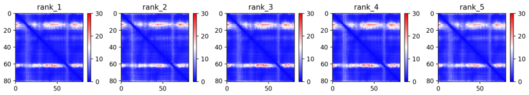

Figure 10: PAE of algo_mut

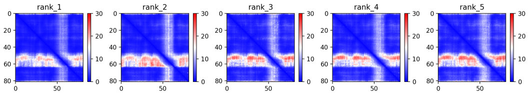

Figure 11: PAE of helix_mut_pro

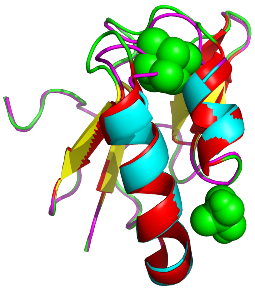

Figure 12: Superimposed structure of wildtype (green, yellow, red) and algo_mut (cyan, red, magenta)

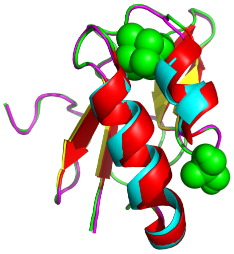

Figure 13: Superimposed structure of wildtype (green, yellow, red) and helix_mut_ala (cyan, magenta, red)

Among the three predicted models, **helix_mut_ala** emerges as the most reliable, with consistently high pLDDT and pTM scores indicating robust local and global structural integrity. In comparison, **algo_mut** and **helix_mut_pro** demonstrate strong local prediction accuracy but slightly reduced global coherence, as evidenced by their moderate pTM scores and higher deviations in PAE plots.

6 Superimposition and structure evaluation

6.0.1 Superimposition and Structural Evaluation

To evaluate the structural deviations between the native wild-type BtFd and its mutant variants, the structures were superimposed using PyMOL. The `align` function in PyMOL was employed, which calculates the Root Mean Square Deviation (RMSD) based on the α-carbon (Cα) atoms of aligned residues. RMSD values provide a quantitative measure of structural differences, offering insight into the structural stability and alterations introduced by the mutations.

Superimposition Results Three mutant structures ("mutant", "helix_mut", and "helix_mut_pro") were aligned to the wild-type BtFd to assess structural conservation and deviations. The results are summarized as follows:

Wild-type vs. Mutant:

- **RMSD**: **0.310 Å** (435 aligned atoms).
- Minimal structural deviations were observed, with slight positional shifts localized near the active site. These results indicate that the mutant structure preserves the overall fold and structural integrity of the wild-type.
- Refer to Figure 12 for the superimposed structures.

Wild-type vs. Helix_mut:

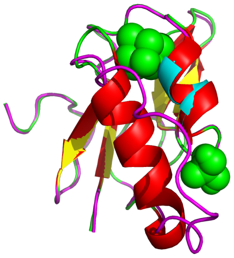

Figure 14: Superimposed structure of wildtype (green, yellow, red) and helix_mut_pro (cyan, magenta, red)

- **RMSD**: **0.457 Å** (473 aligned atoms).
- The helix_mut structure exhibited slightly greater deviations compared to the general mutant, likely due to helical modifications. However, the deviations remain modest, reflecting a well-preserved backbone.
- See Figure 13 for a visual representation.

Wild-type vs. Helix_mut_pro:

- **RMSD**: **0.359 Å** (402 aligned atoms).
- Despite the extensive proline substitutions within the helix, the overall structural deviations remained moderate. The observed RMSD suggests that while secondary structure elements, such as helices, may have been disrupted, the global fold remained intact.
- Refer to Figure 14 for the superimposition.

Discussion of Results The RMSD values derived from the PyMOL alignments provide compelling evidence of structural conservation between the wild-type and its mutants. Specific insights include:

1. **Minimal deviation in the mutant structure**: The mutant exhibited the lowest RMSD (0.310 Å), consistent with its limited modifications localized to the active site. This indicates that the structural integrity of the wild-type is largely maintained.
2. **Localized deviations in helix_mut**: The helix_mut structure demonstrated slightly higher RMSD (0.457 Å), reflecting the impact of targeted helical modifications. Despite this, the backbone remained stable, underscoring the resilience of the overall fold.
3. **Structural resilience in helix_mut_pro**: The helix_mut_pro structure, with significant proline substitutions, retained a relatively low RMSD (0.359 Å). While secondary structural disruptions are expected with proline insertions, the global fold remained robust.

Figures

- **Figure 12**: Superimposition of wild-type BtFd and mutant structure.
- **Figure 13**: Superimposition of wild-type BtFd and helix_mut.
- **Figure 14**: Superimposition of wild-type BtFd and helix_mut_pro.

These results, coupled with the visual representations, highlight the structural conservation of BtFd across its mutants, despite localized deviations. This analysis emphasizes the robustness of the wild-type fold and the specific impact of targeted mutations on structural integrity.

7 Conclusion

This project aimed to explore the structural integrity of Bacillus thermoproteolyticus ferredoxin (BtFd) under conditions relevant to early Earth hydrothermal vent systems. While the initial goal included molecular dynamics (MD) simulations under extreme conditions, such as acidic pH and elevated temperatures, technical challenges and time constraints limited the scope to simulations under neutral pH and room temperature. Despite these limitations, the results obtained provide valuable insights and pave the way for future studies.

Even without MD simulations in extreme conditions, the findings demonstrate the remarkable resilience of BtFd. Notably, the protein retained its structural integrity even after mutating the second alpha helix entirely to alanine. This resilience was unexpected and underscores the adaptability of BtFd's structure. It suggests that ancient ferredoxins, like BtFd, could have played a critical role in early life by remaining stable in dynamic and harsh environments. This observation aligns with the hypothesis that ferredoxins were central to driving prebiotic metabolic processes using gradients in hydrothermal vent systems.

The results also reaffirm the importance of conducting MD simulations under conditions mimicking hydrothermal vents, which would provide a more comprehensive understanding of how such proteins behaved in ancient environments. These simulations are an exciting avenue for future exploration, potentially revealing further details about the role of ferredoxins in the origin of life.

In summary, this study highlights the structural resilience of BtFd and its potential implications for early life evolution. Although the initial scope was constrained, the findings open up intriguing questions about the adaptability of ancient proteins and their importance in prebiotic chemistry. Future work will focus on MD simulations to deepen our understanding of these mechanisms.

Data availability

All ascociated files will be uploaded to the following github repository:
https://github.com/UrbanMidgets/structural_bioinformatics_exam
The script will also be attached in the hand-in of this project.

References

[1] B. Alberts et al. *Molecular Biology of the Cell.* Garland Science, 2014.

[2] S. Ekramullah. An assessment of life. *Structural Bioinformatics Review*, 2023.

[3] S. Ekramullah. On the rtca pathway as a primary carbon fixation pathway at the origin of life. *Origin of Life Studies*, 2023.

[4] Keiichi Fukuyama, Toshihiro Okada, Yoshimitsu Kakuta, and Yasuhiro Takahashi. Atomic resolution structures of oxidized [4fe-4s] ferredoxin from bacillus thermoproteolyticus in two crystal forms: systematic distortion of [4fe-4s] cluster in the protein. *Journal of Molecular Biology*, 315:1155–1166, 2002.

[5] G. Kim, S. Lee, E. L. Karin, H. Kim, Y. Moriwaki, S. Ovchinnikov, M. Steinegger, and M. Mirdita. Easy and accurate protein structure prediction using colabfold. *Nature Protocols*, 2024.

[6] S. Krimm and J. Bandekar. Vibrational spectroscopy and conformation of peptides, polypeptides, and proteins. *Adv. Protein Chem.*, 38:181–364, 1986.

[7] E. Smith and H. J. Morowitz. *The Origin and Nature of Life on Earth.* Cambridge University Press, 2016.

[8] A. Warshel. Electrostatic basis of structure-function correlation in proteins. *Chem. Rev.*, 97:2523–2544, 1997.