Correlation between SARS-CoV-2 Spike Protein Sequence Variants and Structure Variance Using Phylogenetic Tree and Protein Simulation

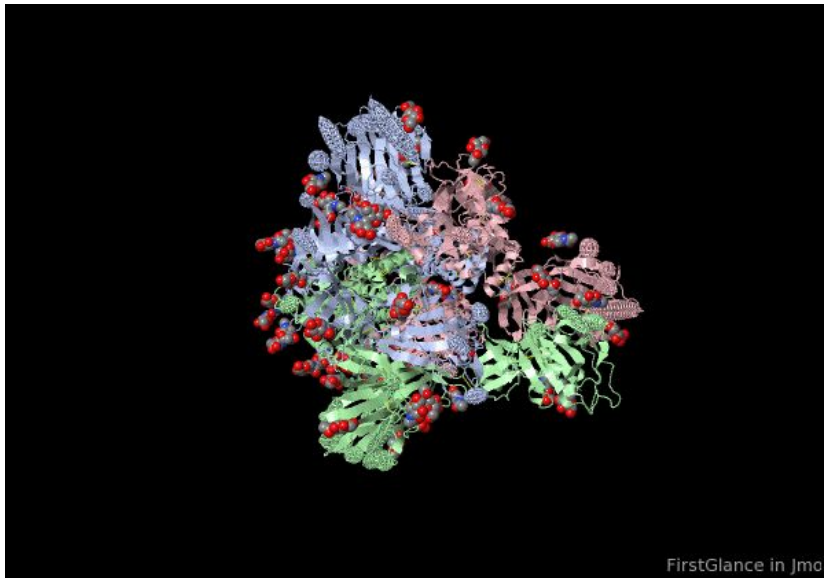
LCM Team: Katrina **L**iu, Kefan **C**ao, Christina **M**a

02251 Great Ideas in Computational Biology, 2021 Fall

May 3rd, 2021

Background

SARS-COV-2 Spike(S) Protein: The S protein binds to the receptor and promotes virus entry into the cell.



Model 6VYB: SARS-CoV-2 spike ectodomain structure (open state)

(DOI:10.1016/j.cell.2020.02.058)

GIF created from https://proteopedia.org/
PDB Structure from Alexandra C. Walls, Young-Jun Park, M. Alejandra Tortorici, Abigail Wall, Andrew T. McGuire, David Veesler, Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein, Cell, Volume 181, Issue 2, 2020, Pages 281-292.e6, ISSN 0092-8674, https://doi.org/10.1016/j.cell.2020.02.058.

Background & Research Question

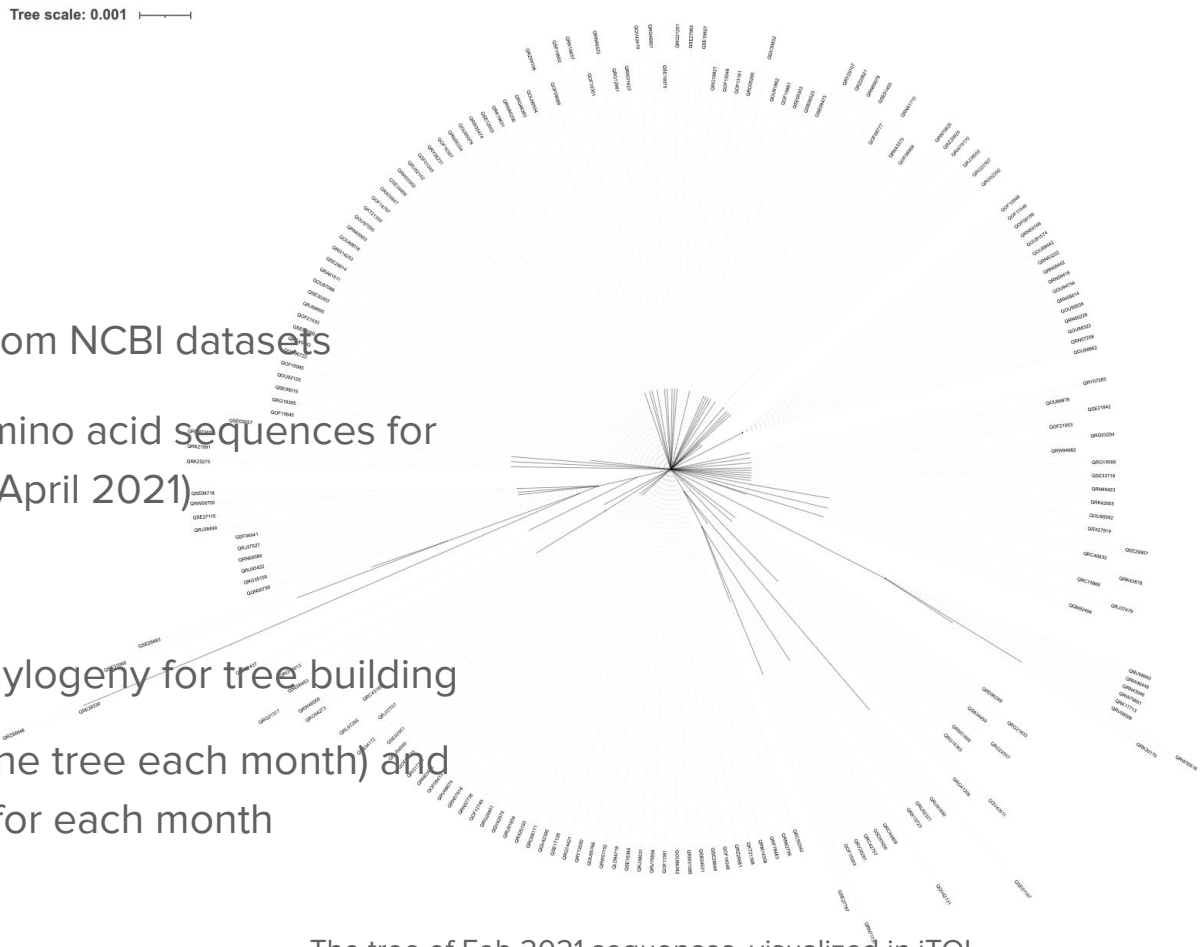
Research Question: Is there a correlation between SARS-CoV-2 spike protein sequence variants and their structural-functional differences?

Hypothesis: For amino acids sequences with larger tree distances, their protein structures should be less similar.

- Protein Variant Similarity:
  - **Patristic Distance**: the sum of the lengths of branches that link two nodes in a tree
- Protein Structure Similarity:
  - **QH**: a metric for structural homology that measures structural conservation
  - **RMSD**: Root Mean Square Deviation
  - **PI**: Percent Identity

Methodology

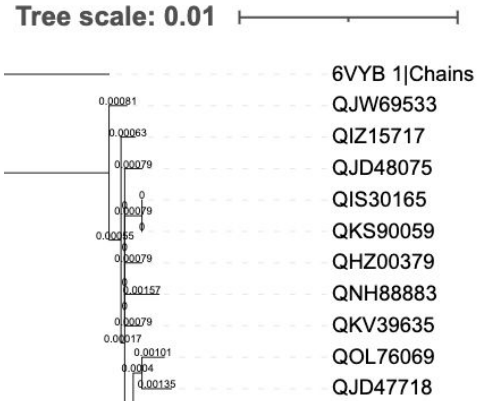
- Preparing Data:
  - Retrieving Sequences from NCBI datasets
  - Randomly select 200 amino acid sequences for each month (Feb 2020-April 2021)
- Trees Building:
  - MUSCLE and Simple Phylogeny for tree building
  - iTOL for visualization (one tree each month) and cherry pick sequences for each month

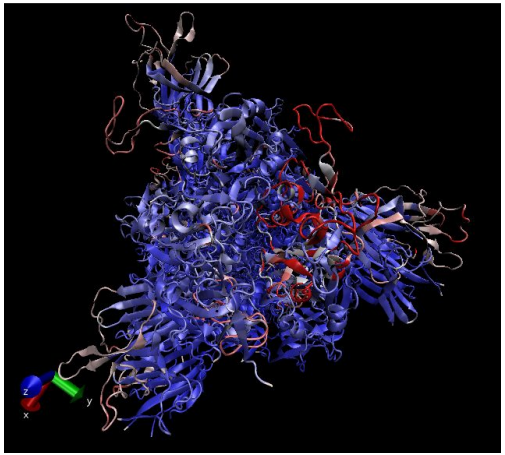


The tree of Feb 2021 sequences, visualized in iTOL

Methodology

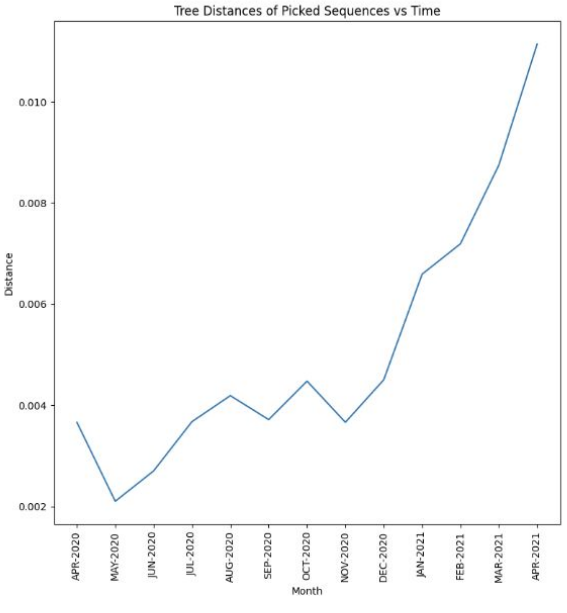
- Tree Distance Calculation: within month & against 6VYB
  - Within months: Patristic Distance calculated using dendropy
  - Against 6VYB: Build tree with picked sequences and rerooted at 6VYB, sum of branch lengths
- Structural Comparison:
  - SWISS-MODEL for homogenous modeling of the picked sequences
  - VMD for visualizing structures and obtaining structural comparison data

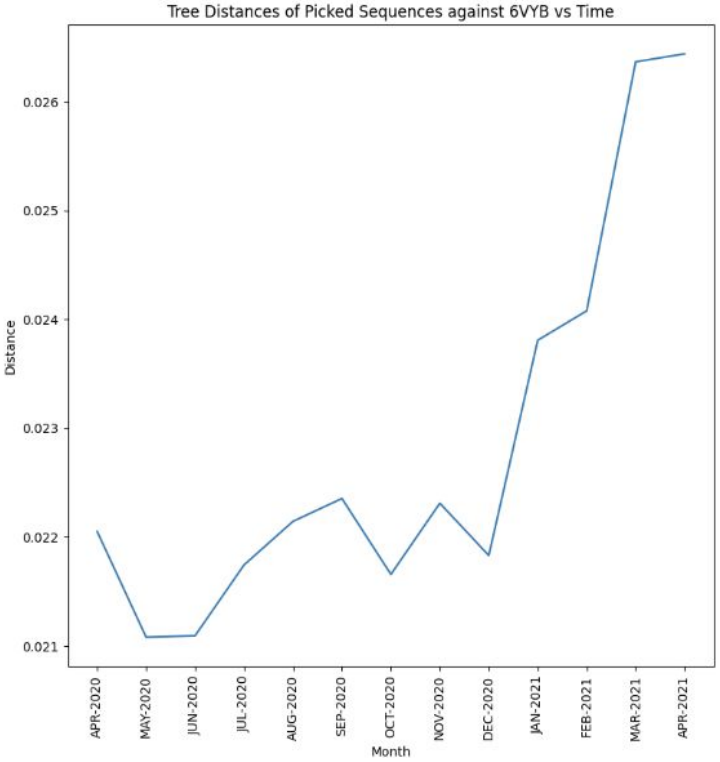




Results

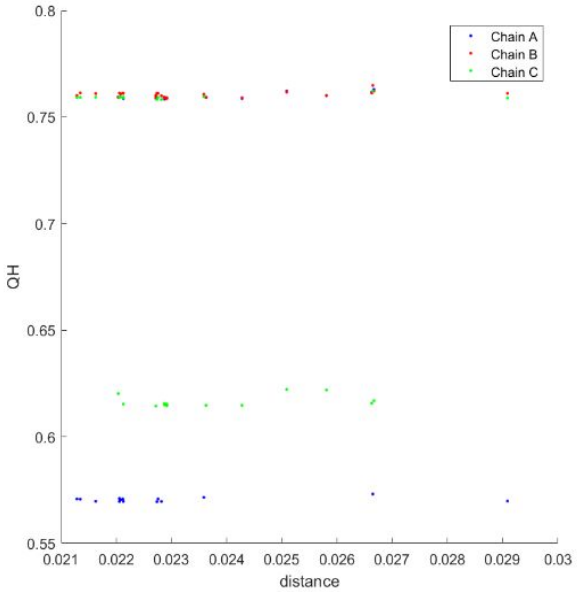
Chronological tree distance within month
& against 6VYB

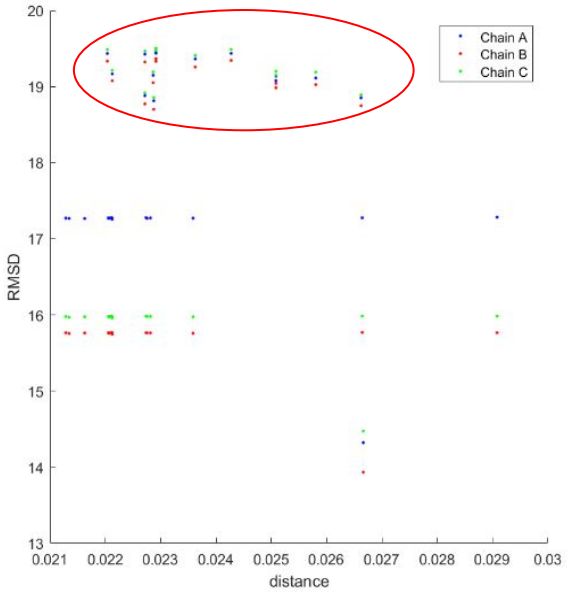


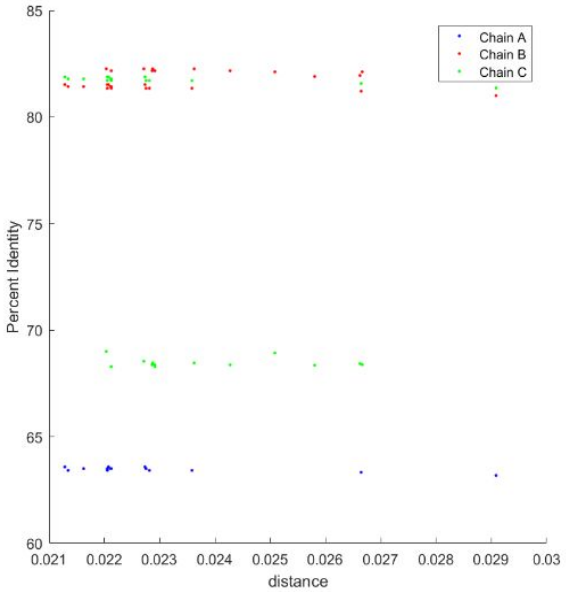


Results

Tree Distance vs. Protein Structure Similarity (QH, RMSD, PI) against 6VYB

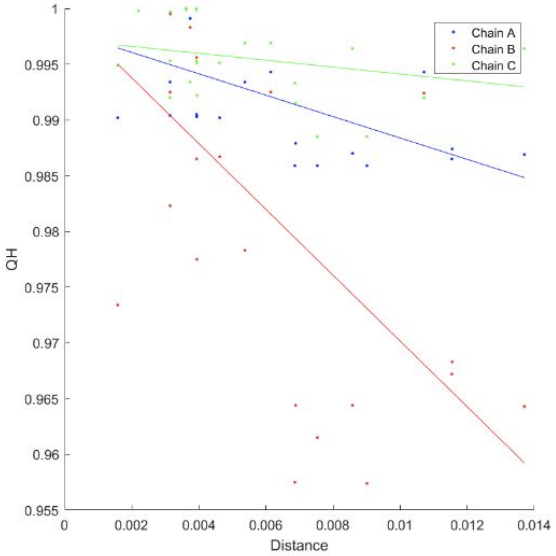


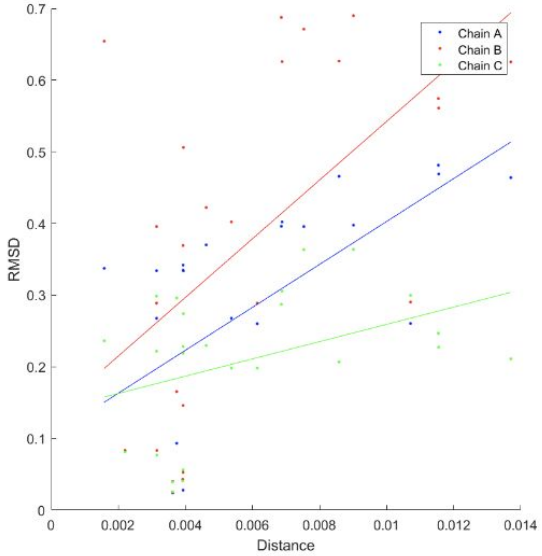


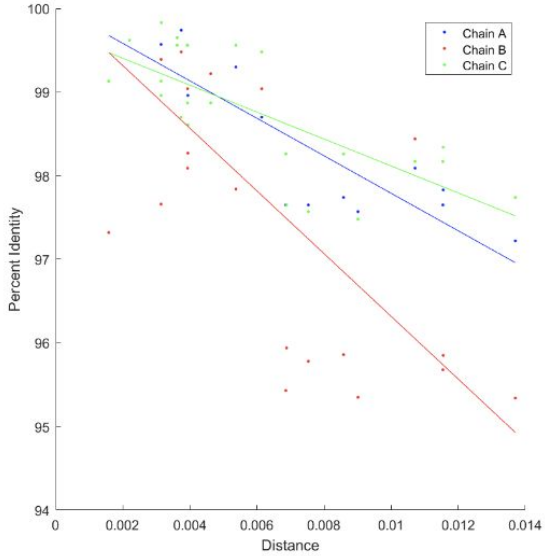


Results

Tree Distance vs. Protein Structure Similarity (QH, RMSD, PI) within month with Regression

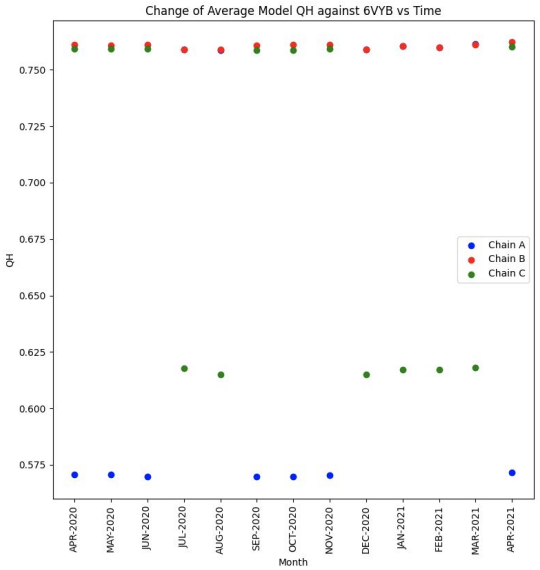


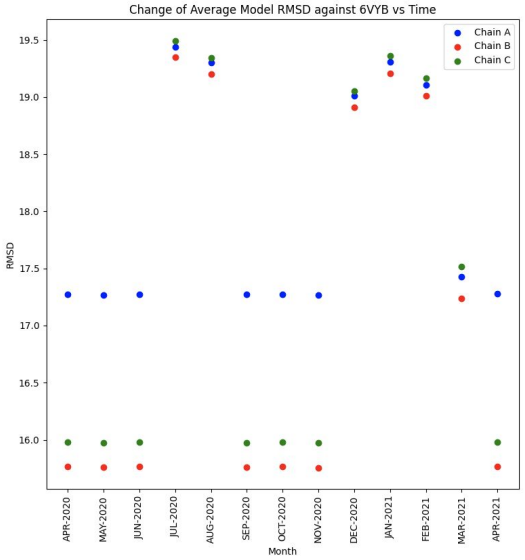


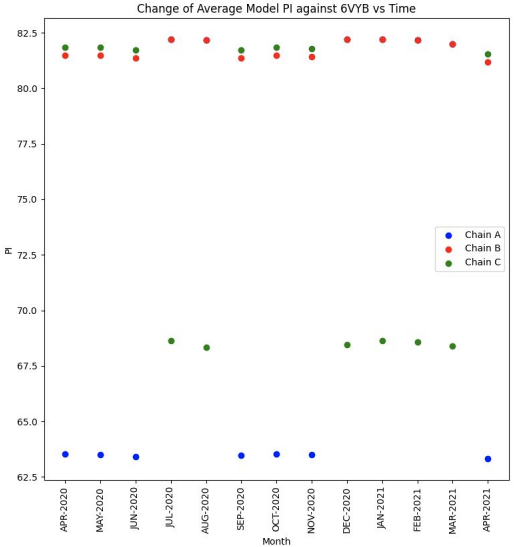


Results

Chronological protein structure similarity (QH, RMSD, PI) against 6VYB

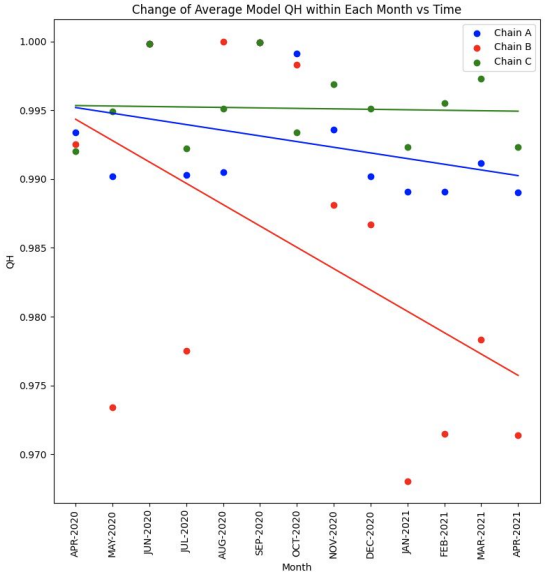


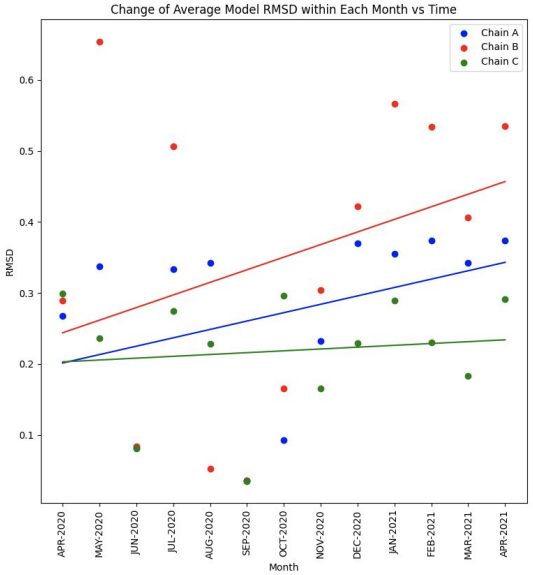


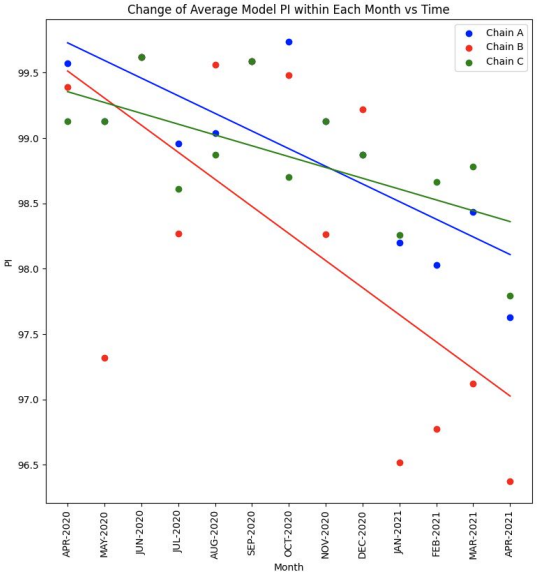


Results

Chronological protein structure similarity (QH, RMSD, PI) within month





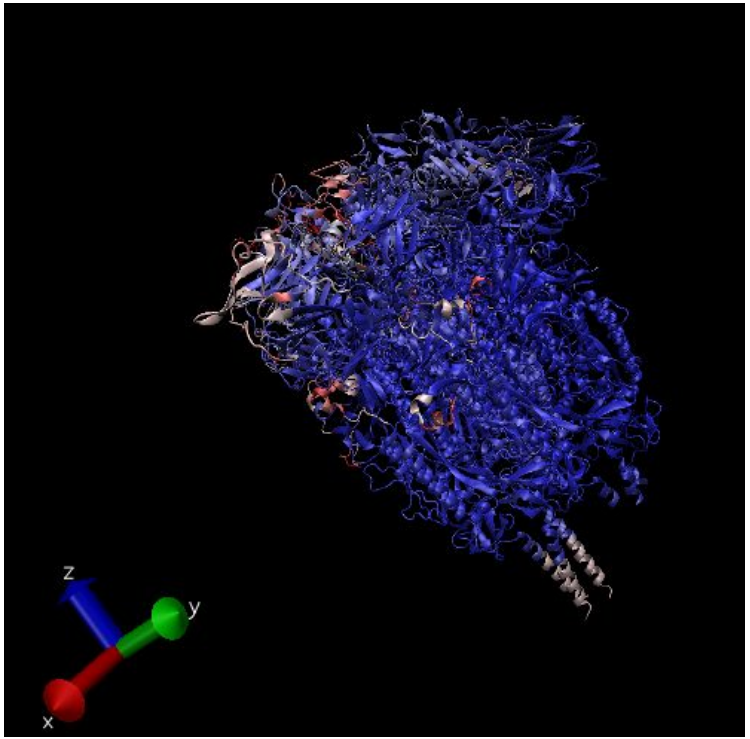


Conclusion

- Structural Similarity vs Tree Distance
  - Against Model 6VYB:
    - The data does not display a significant positive correlation between structural similarity and tree distance, but there are clusterings of similarity regardless of tree distances.
  - Between Models within Each Month:
    - The structural similarity vs. tree distance for each month exhibits negative correlations for QH and PI, and a positive correlation for RMSD.
    - Although the correlations are insignificant, this is still consistent with our hypothesis that the structural differences.of proteins is correlated with their sequence variations.

Reflection

- Abnormality in data
  - Bias from 6VYB
  - Random selection
  - Cherry pick choices
  - Swiss model templates
  - VMD alignment
  - Amount of data points



A flawed protein structural prediction in VMD

Future Improvements and Extensions

- Improvements
  - Sampling more data points
  - Selecting sequences using better criterion
  - Generate the protein structure of 6VYB like other amino acids sequences (Swiss model)
- Extensions
  - A more computational and automated way to collect data and perform analysis.
  - Compare against closed state spike protein model (6VXX)
  - Shorten the time interval for selecting sequences
  - Assumption of clustering of data points implies different types of mutations

Reference for the Tools

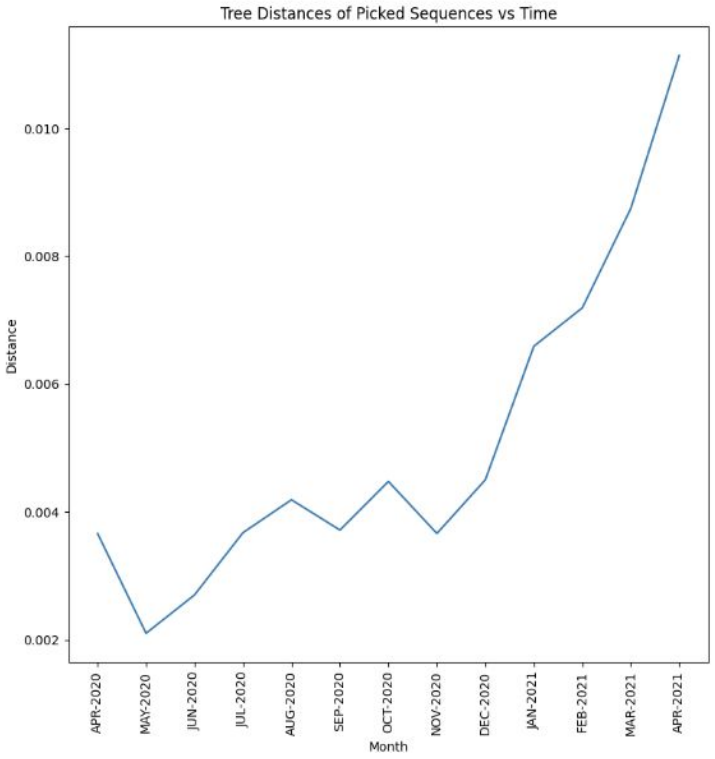
- National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2021 May 01]. Available from: https://www.ncbi.nlm.nih.gov/
- Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019 Jul;47(W1) W636-W641. doi:10.1093/nar/gkz268. PMID: 30976793; PMCID: PMC6602479.
- Letunic, I., & Bork, P. (2021, April 22). Interactive tree of life (itol) v5: An online tool for phylogenetic tree display and annotation. Retrieved May 01, 2021, from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab301/6246398#
- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46, W296-W303 (2018).
- Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38.

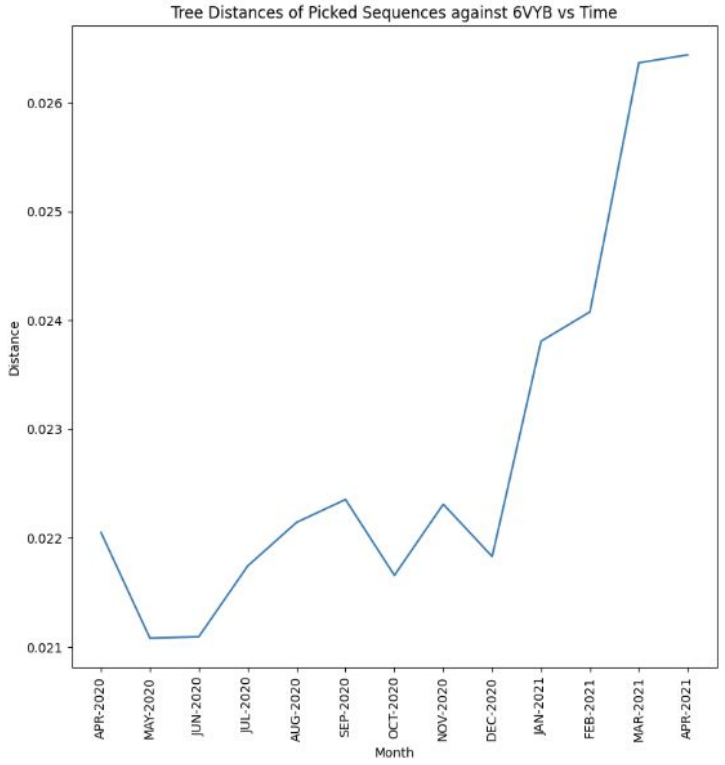
Thank you!

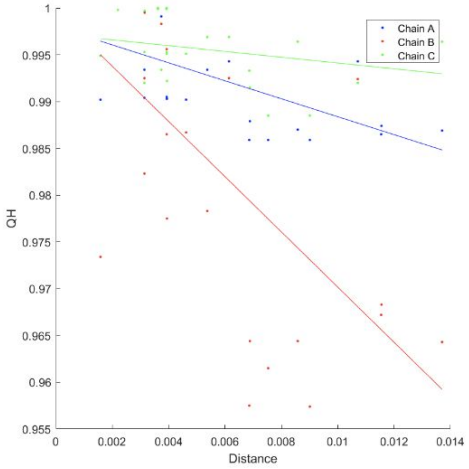
Any question?

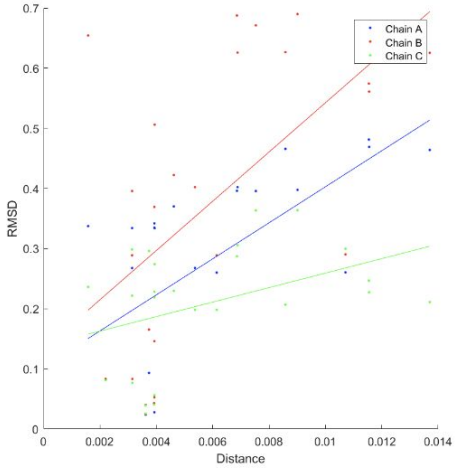
Results

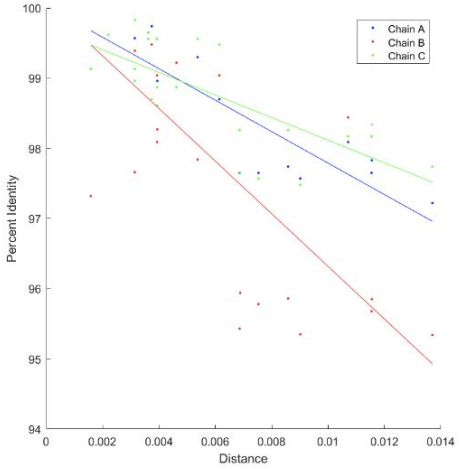
Chronological tree distance within month & against 6VYB

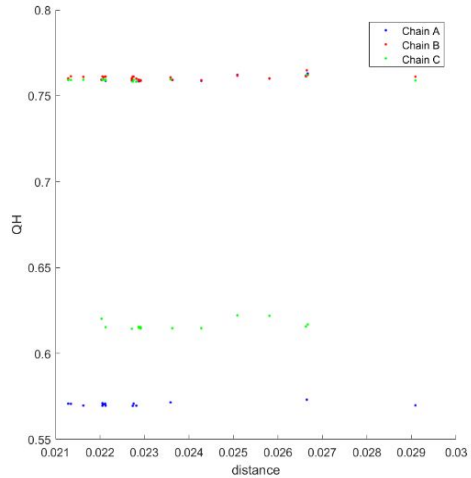


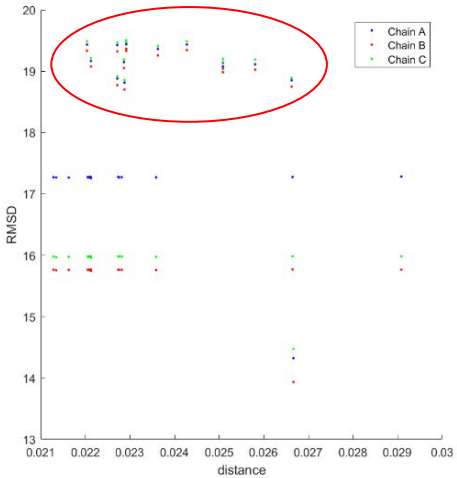


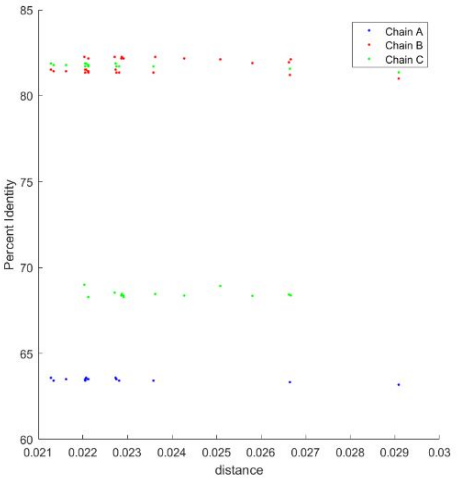

Chain A
Chain B
Chain C
QH
Distance


Chain A
Chain B
Chain C
RMSD
Distance


Chain A
Chain B
Chain C
Percent Identity
Distance

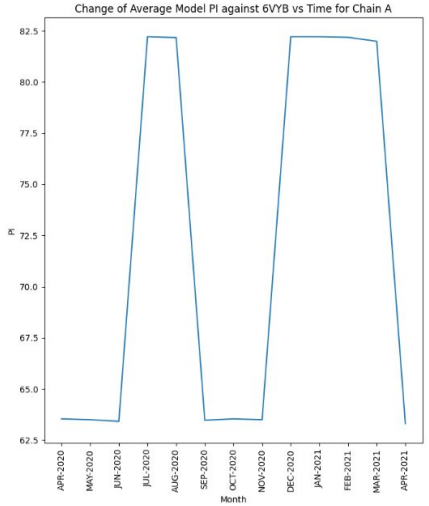

Chain A
Chain B
Chain C
QH
distance

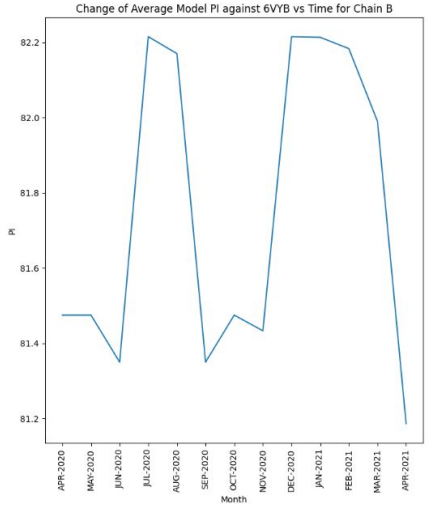

Chain A
Chain B
Chain C
RMSD
distance

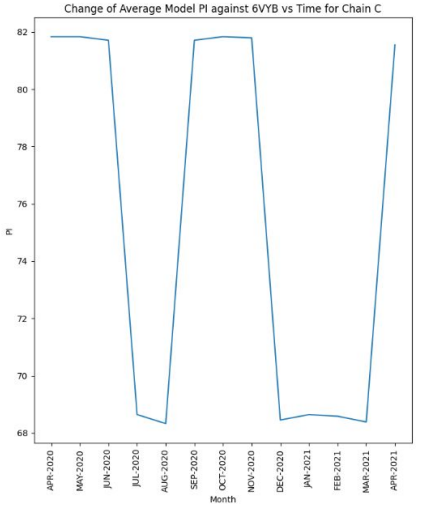

Chain A
Chain B
Chain C
Percent Identity
distance

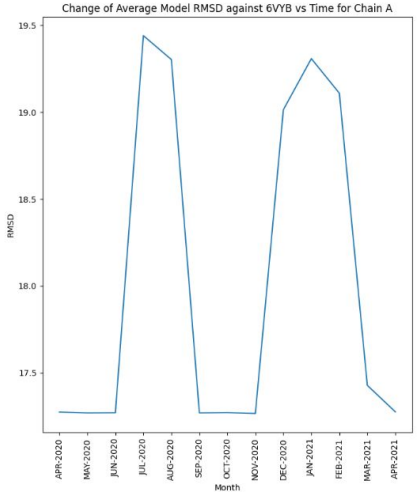
Results

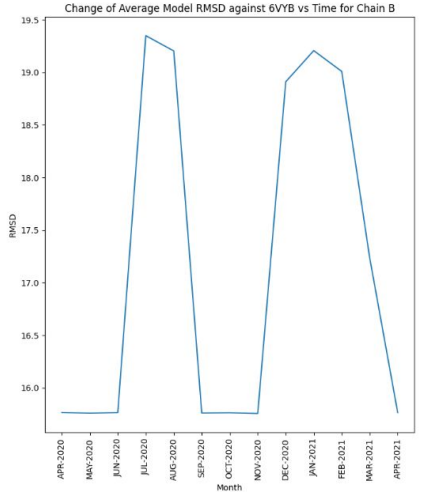
Chronological structural similarity (PI & RMSD) by Chain A/B/C

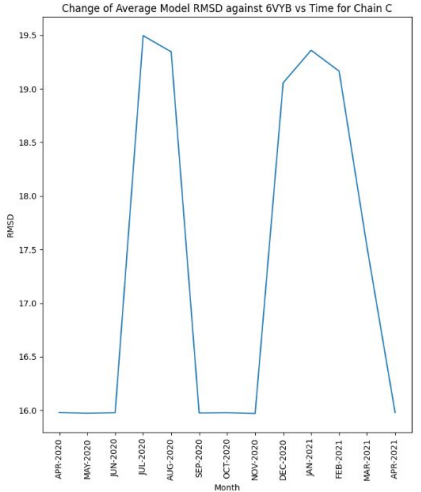


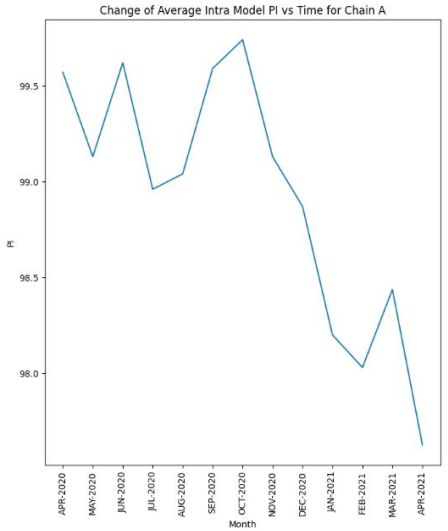


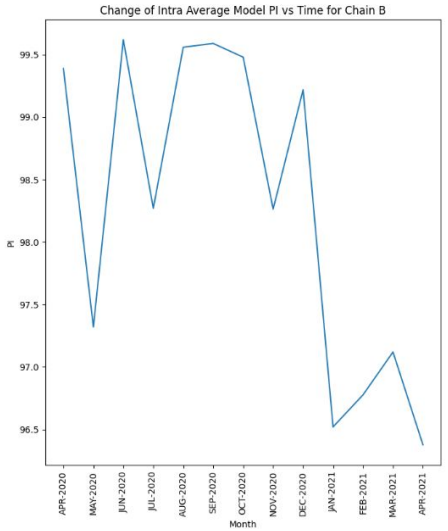


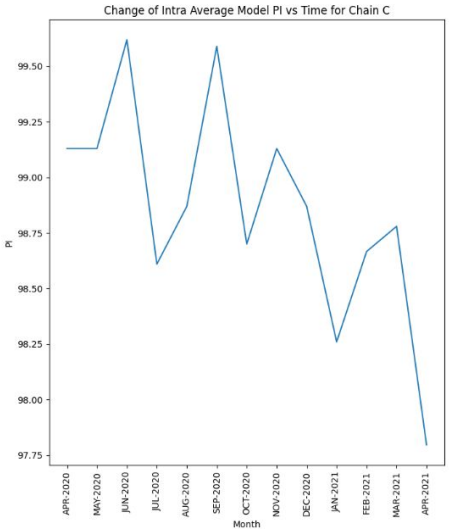


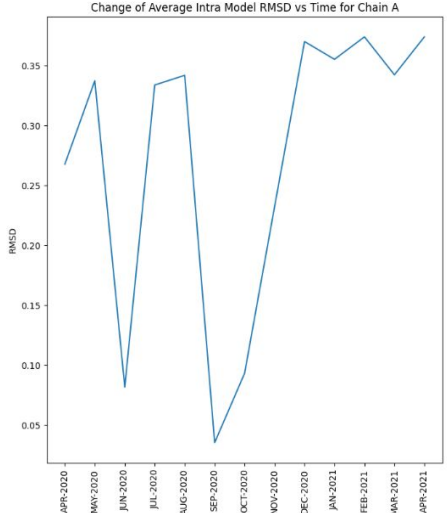


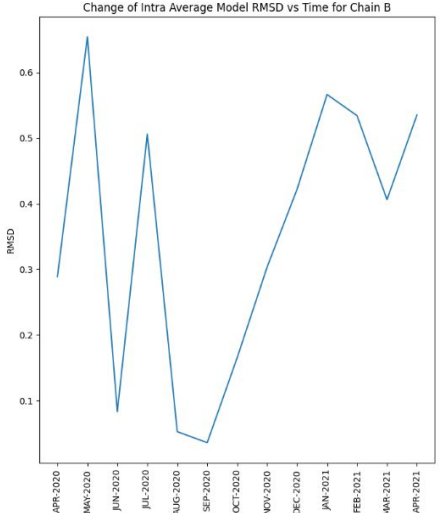


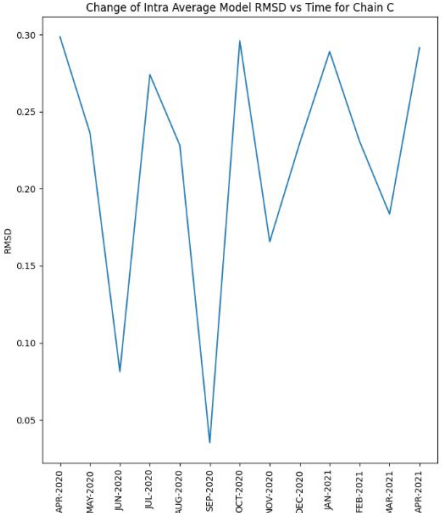

Change of Average Intra Model PI vs Time for Chain A
PI
Month

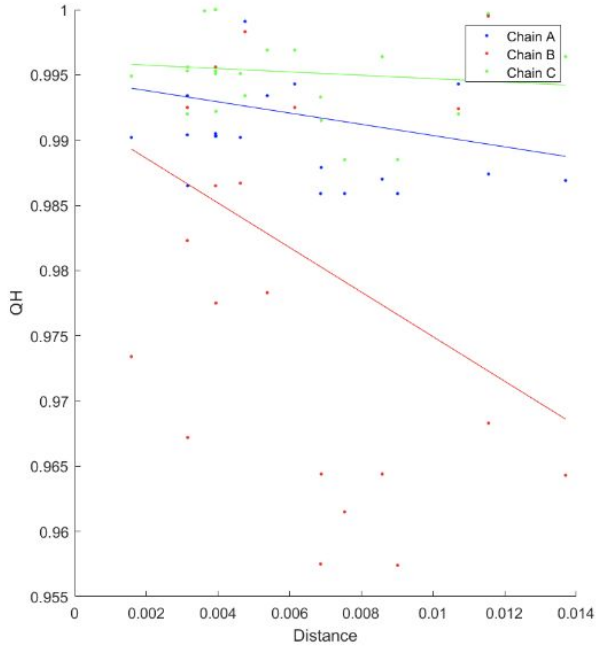

Change of Intra Average Model PI vs Time for Chain B
PI
Month

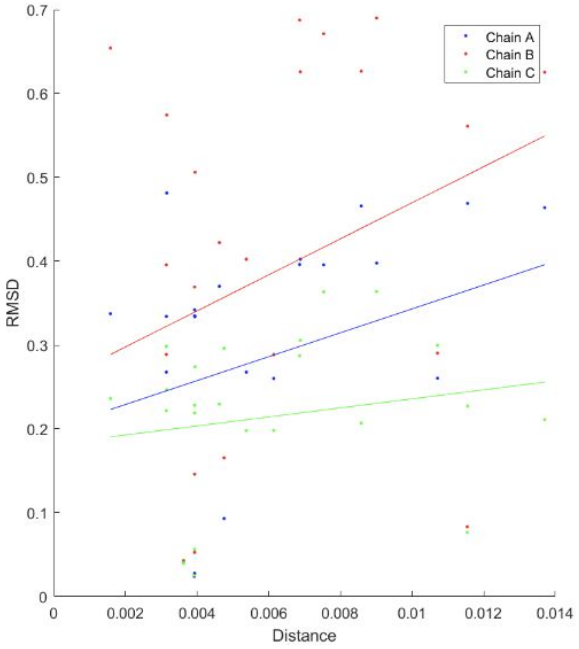

Change of Intra Average Model PI vs Time for Chain C
PI
Month

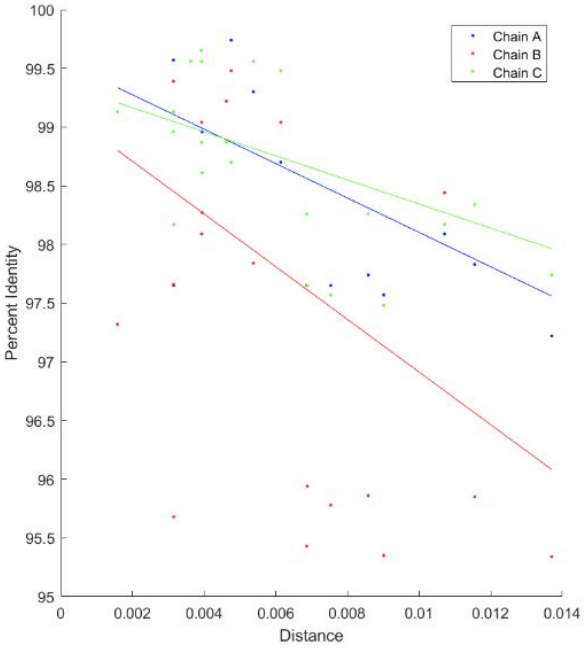

Change of Average Intra Model RMSD vs Time for Chain A
RMSD


Change of Intra Average Model RMSD vs Time for Chain B
RMSD


Change of Intra Average Model RMSD vs Time for Chain C
RMSD


Chain A
Chain B
Chain C
QH
Distance
1
0.995
0.99
0.985
0.98
0.975
0.97
0.965
0.96
0.955
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014


Chain A
Chain B
Chain C
RMSD
Distance
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014


Chain A
Chain B
Chain C
Percent Identity
Distance
100
99.5
99
98.5
98
97.5
97
96.5
96
95.5
95
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014