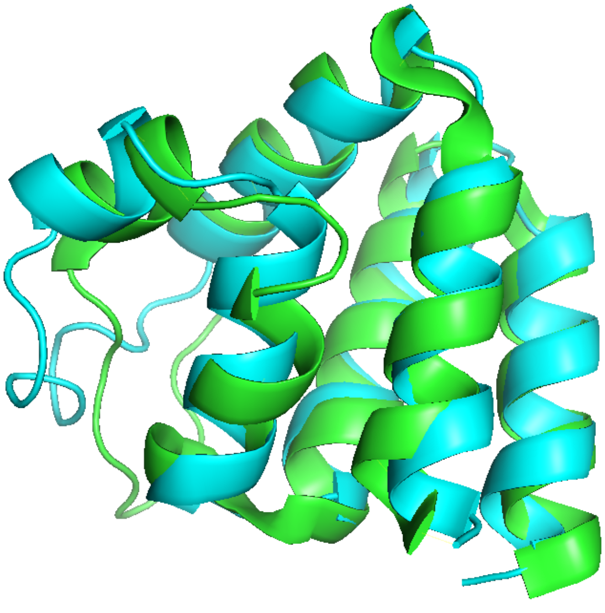
Evaluating the Efficiency of Rosetta Ab Initio and Numerical Linear Algebra in Protein Structure Prediction



Spyros Alvanakis
Konstantinos Giatras

Final Project Report for the Algorithms in Structural Bioinformatics Course

MSc DSIT 2022-2023

Contents

Abstract

This project presents a comparative study of two computational methods for protein structure prediction - the Rosetta Ab Initio and a Linear Algebra Algorithm - critical for advancements in drug discovery and understanding diseases at a molecular level. The Rosetta Ab Initio method uses a Monte Carlo optimization technique to predict protein tertiary structures from amino acid sequences alone, while the Linear Algebra Algorithm utilizes NMR spectroscopy data in the form of a Cayley-Menger matrix to generate protein conformations. The efficacy of these methods was evaluated using five proteins of varying sizes. The Rosetta Ab Initio showed a positive correlation between the number of predicted structures and the accuracy of predictions, albeit less effective for larger proteins. Conversely, the Linear Algebra Algorithm was inefficient for larger proteins due to high computational complexity, with results exhibiting higher than expected RMSD values. It was concluded that these methods have their individual limitations, making them non-combinable and subject to choice depending on the protein size and available computational resources. Despite these constraints, the potential for improving these methods, potentially with machine learning and computational science advancements, is highlighted. The challenge of high accuracy prediction for larger proteins is noted as a necessary area for future research.

1 Introduction

Understanding the structure of proteins is crucial in many fields, particularly in drug design and understanding diseases at a molecular level. The structure of a protein determines its function, so predicting protein structure from sequence (a problem known as protein folding) is a key problem in biology.

Computational methods that assist in predicting protein structure provide an alternative to experimental methods like X-ray crystallography and cryo-electron microscopy, that can be time-consuming and expensive. By comparing and improving these computational methods, we can potentially speed up the process of protein structure determination, enabling faster drug discovery and disease understanding.

For the purposes of our project, we searched for well-documented, publicly available computational methods with 3D structure prediction capabilities. We chose to evaluate and compare the efficiency of two such methods:

- **Rosetta Ab Initio**[14], part of the widely used Rosetta software suite (includes algorithms for computational modeling and analysis of protein structures), which uses a Monte Carlo optimization algorithm and energy minimization techniques to predict three-dimensional protein structures from amino acid sequences, generating a multitude of predicted protein structures.

- A **Linear Algebra Algorithm**, developed and described in the 2005 paper "Molecular Conformation Search by Distance Matrix Perturbations"[10], that perturbs an input Cayley-Menger matrix, iteratively reducing its rank to at most five using singular value decomposition and gradient-based approaches, with the goal of generating a matrix that can be used to extract 3D coordinates of backbone atoms from real or simulated NMR (Nuclear Magnetic Resonance) distance data.

2 Algorithms

2.1 Rosetta Ab Initio

Ab initio (Latin for "from the beginning") structure prediction refers to the method of predicting a protein's tertiary structure from its amino acid sequence alone, with no reference to known structures. Ab initio algorithms, such as Rosetta's, are used to achieve this task[12]. Rosetta Ab Initio[14][3] accepts a protein sequence in FASTA format as its primary input, as well as a secondary structure file (ss2), a 3mer-fragment and a 9mer-fragment file, which are provided to guide the fragment assembly process[1]. Additionally, if the user desires to obtain the root mean square deviation (RMSD) score of the predicted structures, they have the option to include the pdb file of the protein as part of the input.

The secondary structure file employs predictions to classify each amino acid in the sequence as either an alpha helix, beta pleated sheet, or loop region, each classification coming with a certain probability. During the fragment assembly process, the 3mer-fragment and 9mer-fragment files are utilized. These are produced based on resemblances to established protein structures and play a crucial role in guiding the quest for the most probable protein configurations. These fragments are selected with the goal of portraying local structures, such as helices or beta strands, that are expected to feature in the final protein structure. The fragment libraries, holding hundreds of thousands of brief amino acid sequences, each associated with a unique secondary structure, are key to this process. These fragments are put together to simulate the secondary and tertiary structures of the protein, with the algorithm exploring different combinations to investigate the conformational space. This "fragment library" strategy effectively narrows the search space, thereby enhancing the efficiency of the prediction process[14].

During the ab initio folding procedure, Monte Carlo simulations are utilized to simulate protein folding and to refine the structure based on the input data. The Monte Carlo method, a type of randomized algorithm, is often used to sample conformational space in such predictions, because the conformational space of proteins is vast, and deterministic algorithms would be too computationally expensive. It makes random, incremental modifications to the protein's structure and estimates the energy of each new conformation via a scoring function. If the new conformation has lower energy (indicating a more stable structure), it's accepted. If it results in a higher energy structure, the last movement is reverted and it moves onto another amino acid (it may still be accepted with a certain probability, to avoid getting stuck in local minima)[14].

The final output of the ab initio algorithm is a collection of potential protein conformations represented by the generated structure models. The selection of native-like models from the pool of predicted structures can be performed using various criteria, such as energy functions or scoring functions that assess the quality and fitness of the models.

Rosetta plays an important role in protein structure prediction and design, offering a powerful computational tool that enables the exploration of protein folding, protein-protein interactions, and protein-ligand interactions, thereby advancing our understanding of fundamental biological processes and facilitating the development of novel therapeutics and biomaterials.

2.2 Linear Algebra Algorithm

In NMR spectroscopy, the basic information we get is the frequency and intensity of nuclear spin transitions, which can be influenced by the structure of the molecule and the magnetic properties of the atoms. However, noise in the measurement can make this data difficult to interpret directly. Linear algebra can be utilized to process and interpret noisy NMR data, including the computation of a distance matrix.

This linear algebra algorithm's main function (`mconf`) attempts to take as input a Cayley-Menger matrix (derived from real or simulated NMR distance data), a tolerance level, and a flag. It first checks if the input matrix is symmetric or not. Symmetry is tested by subtracting the matrix from its transpose; if the result is a zero matrix, the input matrix is symmetric. If it is not symmetric (like in our case), it is treated as a bounds matrix. The function generates a set of perturbation entries with the `perbasis` function. After that, it calls the `bnd2mid` function to convert the bounds matrix to a symmetric matrix using the flag variable. The algorithm then calculates the first 6 singular values of the input matrix using the `svds` function.

The main loop begins where the function aims to perturb the input matrix such that the sixth singular value becomes less than the tolerance value. This is done using the `svred` function which takes the input matrix, the perturbation entries, the tolerance level, and the bounds matrix as inputs and returns the perturbed matrix, the new sixth singular value, and the iteration count. If at any point, the new sixth singular value is greater than the tolerance or has increased compared to the initial sixth singular value, the function breaks the loop and proceeds to the next step. It then checks if any of the entries in the new matrix violate the bounds specified in the bounds matrix, and reports these violations.

Finally, the algorithm outputs the new conformation in the form of a symmetric Cayley-Menger matrix. The conformation in this context refers to an acceptable solution or state that the molecule can assume based on the given matrix (embeddable in 3D space)[10].

Linear algebra is critical in protein structure prediction due to its capacity to efficiently handle high-dimensional and noisy data. It offers a robust mathematical basis for modeling complex patterns and dependencies in protein structures, enhancing prediction accuracy. Its application is vital for advancements in bioinformatics, drug discovery, and understanding molecular disease mechanisms.

3 Data and Methods

Our workflow (Figure 1) involved familiarizing ourselves with and executing the aforementioned algorithms, comparing their results, and, time permitting, leveraging the outcomes of the Rosetta Ab Initio algorithm in combination with the Linear Algebra algorithm. The goal was to investigate if we could obtain a conformation with an even smaller RMSD compared to the ground truth protein.

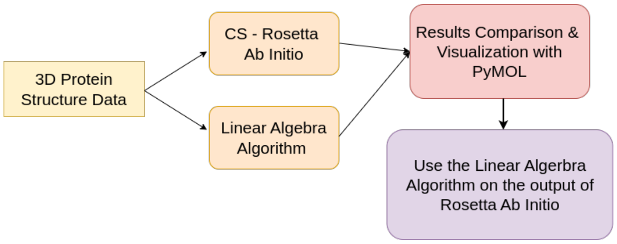

3D Protein Structure Data
CS - Rosetta Ab Initio
Linear Algebra Algorithm
Results Comparison & Visualization with PyMOL
Use the Linear Algerbra Algorithm on the output of Rosetta Ab Initio

Figure 1: Workflow Diagram

3.1 Rosetta Ab Initio

We obtained the Rosetta software from the official website at `https://www.rosettacommons.org/software/`. To ensure proper access to the software, we also acquired the necessary license.

To employ the Rosetta Ab Inition algorithm, we selected five different proteins with varying numbers of amino acids to investigate potential differences in the results. The protein structures 2JYV (Human Granulin F)[9], 2M0D (Miz-1 zinc finger 5)[8], 2MPC (Pyrin domain of human Pyrin)[15], 2MS8 (Mitochondrial antiviral-signaling protein)[13], and 6MWM (Bat coronavirus HKU4 SUD-C)[6] were obtained from the Protein Data Bank (`https://www.rcsb.org/`). It is important to note that these protein structures were derived from solution NMR experiments, resulting in pdb files containing protein ensembles (multiple states) for each protein. The number of amino acids and the ensemble states are shown in Table 1.

PDB ID	Num of aa	Num of States
2JYV	32	10
2M0D	30	20
2MPC	90	10
2MS8	102	20
6MWM	81	20

Table 1: Proteins Description

After identifying the proteins and downloading their PDB files, we utilized the Robetta platform to obtain the corresponding fasta files, 3mer and 9mer fragments, as well as the secondary structure prediction files for each protein. Initially we extracted the fasta file using the command `save /path/to/save/protein_name.fasta`. Then, using the fasta files we extracted the rest of the files from `http://robetta.bakerlab.org/fragmentsubmit.jsp`. Subsequently we created specific flags files for each protein and the desired protein conformations. These flags files are essential as they provide Rosetta with the necessary instructions and commands to execute the desired options for the protein structure predictions. An example of the command to run Rosetta Ab Initio is `/path/to/Rosetta/main/source/bin/AbinitioRelax.default.linuxgccrelease @flags`. This command should be executed from the directory where the necessary foldersand flags file are located. An example of a flags file is illustrated in Table 2.

-database	/path/to/Rosetta/main/database
-in:file:native	./2m0d.pdb
-in:file:fasta	./2m0d.fasta
-in:file:frag3	./aat000_03_05.200_v1_3
-in:file:frag9	./aat000_09_05.200_v1_3
-psipred_ss2	./t000_.psipred_ss2
-nstruct	1000
-abinitio::relax	false
-use_filters	true
-abinitio::increase_cyrcles	10
-abinitio::rg_reweight	0.5
-abinitio::RMSD_wt _helix	0.5
-abinitio::RMSD_wt_loop	0.5
-relax::fast	
-out:file:silent	./2m0d_silent_1000_tries.out

Table 2: Flags file example for 2M0D protein, 1000 predicted structures run

The options for the flags file mentioned above were determined based on input from three different sources [5][4][1]. The first six options are self-explanatory, including the path to the Rosetta database and the pdb file of the desired protein, which is an optional input used when the desired outcome includes an RMSD value. It is worth noting that omitting the pdb file significantly reduces the computational time, at the cost of excluding the intrinsic RMSD calculation. For instance, on a local laptop with Linux and a 3rd generation Ryzen 5 processor, the computations for 1000 conformations took approximately 20 hours with the pdb file, compared to 5 hours without the pdb file.

The flags file's remaining parameters include the 3mer and 9mer fragment files, the fasta file, and the secondary structure prediction file. The seventh parameter lets you choose the desired number of protein conformations for examination, typically picking quantities like 10, 100, or 1000, in order to investigate variations in the mean accuracy of the most accurate predicted proteins against the actual protein, utilizing the RMSD value for assessment. The ability to produce predicted structures using Rosetta Ab Initio chiefly relies on the local machine's computational capacity and the size and complexity of the protein being modeled. For an average-sized protein (around 150 residues), a typical run on a standard local machine may generate about 1,000 predicted structures, thus adequately sampling the protein's potential structural diversity. Notably, producing more protein structures does not necessarily assure better outcomes, given factors like the precision of the scoring function and the extent of the conformational search. Hence, the primary objective should be to identify a small set of high-quality models that accurately mirror the protein's native conformation, necessitating a meticulous analysis and selection of the generated models. The remaining parameters generally maintain their default values, with the exception of the final option that produces a silent file containing all the results in a format easy to manage for additional analysis[14].

For visualizing the outputs of Rosetta Ab Initio, we utilized PyMOL[2], a commonly used tool in this research field. Initially, we adopted the root mean square deviation (RMSD) value from Rosetta as the metric for evaluating the fitting of the predicted structures to the ground truth. However, we observed that the RMSD values obtained from Rosetta and PyMOL differed. Specifically, Rosetta tended to calculate higher RMSD values compared to PyMOL. This discrepancy may arise from the different alignment methods employed by these two algorithms. Due to the lack of transparency regarding the RMSD calculations in each algorithm, we decided to independently calculate the RMSD for each predicted conformation of the Rosetta algorithm using our own method.

We extracted the PDB files for the predicted conformations of every protein using terminal commands. In total, we produced 1110 conformations per protein, in three separate runs (10, 100, 1000). Due to the large number of PDB files, we renamed them using terminal commands to make them more manageable in the Python code. The new naming convention followed the format: (Name of protein)_(Number of conformations)_(ID of conformation).pdb.

As described in `Supplementary_Code_1.ipynb` (see attached), for each conformation, we extracted the 'CA' atoms to calculate the RMSD values. To perform the alignment and RMSD calculation, we utilized the SVD.imposer package from BioPython, which aligns the molecules based on their SVD values.

One challenge we encountered was dealing with the significant number of ensemble states in the initial PDB structures (ground truth). To address this, since Rosetta Ab Initio was employed to generate 10, 100, and 1000 predicted structures for every protein, the CA RMSDs were calculated between each predicted structure and a specific protein state, which then computed the mean RMSD for that state. This was done for every state of the protein, helping to pinpoint the minimum mean RMSD and the best corresponding protein state. We regarded the mean RMSD of a predicted conformation as the final RMSD result for that conformation, averaging the RMSD values across all states, and kept track of the smallest RMSD value among the predicted conformations for a given state of the ground truth structures. Then, the predicted structure with the minimum RMSD relative to the best protein state was identified. This entire process was executed for every protein and for each run of 10, 100, and 1000 predicted conformations. Finally, for each protein and each number of predicted conformations, the top 10 RMSD scores were selected along with the corresponding minimum RMSD value, and we documented the state of the ground truth conformation that resulted in the minimum RMSD.

3.2 Linear Algebra Algorithm

The files necessary to run the Matlab code were provided to us as part of the Algorithms in Structural Bioinformatics course and the code was run on Matlab locally.

Since the expected input of the algorithm is a Cayley-Menger matrix derived from real or simulated NMR distance data, we created a script (see `Supplementary_Code_2.ipynb`) that generates a distance matrix using the backbone atom coordinates from the best state of the ensemble of each of the previously used proteins. It then produces a perturbation matrix derived from the distance matrix by introducing noise of up to 2% to each distance value, thereby creating intervals that simulate NMR data. In this perturbation matrix, the upper triangular portion contains the upper boundaries, while the lower triangular portion houses the lower boundaries of these intervals. Finally, the script constructs a Cayley-Menger matrix from the perturbation matrix and exports it in the form of a Matlab file.

To run the code, all the functions and the input files need to be in the same directory, which is opened by Matlab. However, when we attempted to run it using our smallest protein (2JYV) as input, the runtime of the code execution was immensely long. After analysing the code, we discovered that this was a result of its complexity, which predominantly falls within the realm of $\texttt{O(n}^2\texttt{)}$ for most functions, signifying a direct proportionality between the number of operations and the square of the input size. However, the `svred` function, responsible for calculating the Singular Value Decomposition (SVD), exhibits a higher complexity of $\texttt{O(n}^3\texttt{)}$, meaning the operations increase cubically with the input size. In practical terms, when applied to our smallest protein 2JYV, which comprises 320 atoms, this led to the computation of a Cayley-Menger matrix of dimensions 321x321. Consequently, it required approximately 33 million operations solely for the SVD calculation, highlighting the complexity involved in such processes.

Since Rosetta Ab Initio algorithm cannot handle proteins with less than 30 amino acids, because the Robetta server cannot generate the needed 3mer, 9mer and ss2 corresponding files, and taking into account that the Linear Algebra method seems to be unoptimized to handle larger proteins, due to its extreme runtime, we can already conclude that these two methods are, for the most part, not comparable and cannot be used it tandem for further improvement of each other's results.

However, we can still evaluate the efficiency of the Linear Algebra algorithm on its own. For that purpose, we searched for a couple of proteins with ≤ 40 CA atoms. We chose the structures 1ADX (Fifth EGF-like Domain of Thrombomodulin)[7] and 1ANP (Atrial Natriuretic Peptide)[11], which we obtained from the Protein Data Bank and are also ensembles produced by solution NMR experiments. The number of CA atoms and the ensemble states are shown in Table 3.

PDB ID	Num of CA Atoms	Num of States
1ADX	40	14
1ANP	28	11

Table 3: Proteins Description

We used PyMOL to save a single random state from each protein ensemble, by typing `cmd.save("output.pdb", state=number)` in the PyMOL command line interface. After generating the corresponding asymmetric Cayley-Menger matrices in Matlab format for both proteins (we chose a random state for each of the ensembles) using our aforementioned script, we add the files to the directory of the Matlab code. For each protein, we opened the Cayley-Menger matrix, we ran it and it got saved in a default `ans` matrix variable. Then, we ran the code, by typing `mconf(ans,1e-8,1)` in the Matlab command line. After the code ran successfully and the new Cayley-Menger matrix is saved on the same `ans` variable, we check its rank using the `rank(ans)` command and save it as a csv file, using the `csvwrite("name.csv",ans)` command.

At this stage, we enriched our previous Python script (see `Supplementary_Code_2.ipynb`), so that it reads the new Cayley-Menger matrix from the csv file and converts it into a numerical format (floating-point numbers). The rank of this matrix is computed and printed again for confirmation, which indicates its dimensionality (a rank of 5 implies that the protein structure is embeddable in 3D space). The Cayley-Menger matrix is then converted back into a distance matrix, which is further transformed into 3D coordinates using Multidimensional Scaling (MDS). MDS is a technique often used in structural biology to convert distance matrices into Cartesian coordinates. Finally, this script calculates the root-mean-square deviation (RMSD) between the coordinates derived from the original state of the protein (from the pdb file - ground truth) and the new predicted coordinates. It even includes an alignment step before the calculation of RMSD, using the SVD.imposer class from the Bio.PDB module, to ensure the comparison is fair and meaningful. The final RMSD value is then printed out, providing a quantitative measure of the structural difference between the two protein structures.

4 Results

4.1 Rosetta Ab Initio

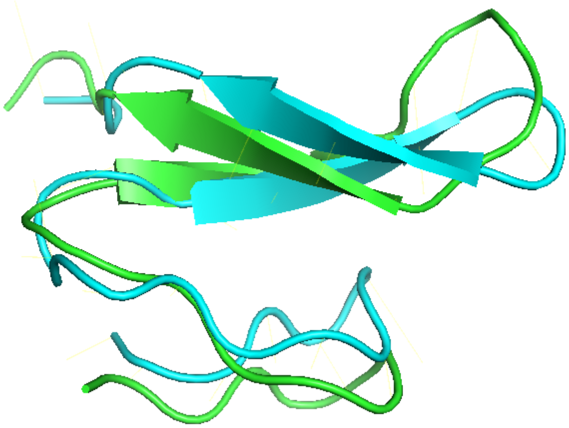


Figure 2: The best predicted conformation of 2JYV with the 4th state of the ground truth coordinates of 2JYV. RMSD = 2.89 Å

The Rosetta algorithm was utilized to create a series of predicted conformations for every protein in the dataset. Each protein yielded 1110 conformations and the resulting PDB files were organized and renamed for efficient management. 'CA' atoms from each configuration were isolated, with alignment and RMSD values evaluated using the SVD.imposer package from Biopython. Notably, RMSD values obtained from the Rosetta algorithm were slightly higher than those from PyMOL, potentially a result of divergent alignment techniques.

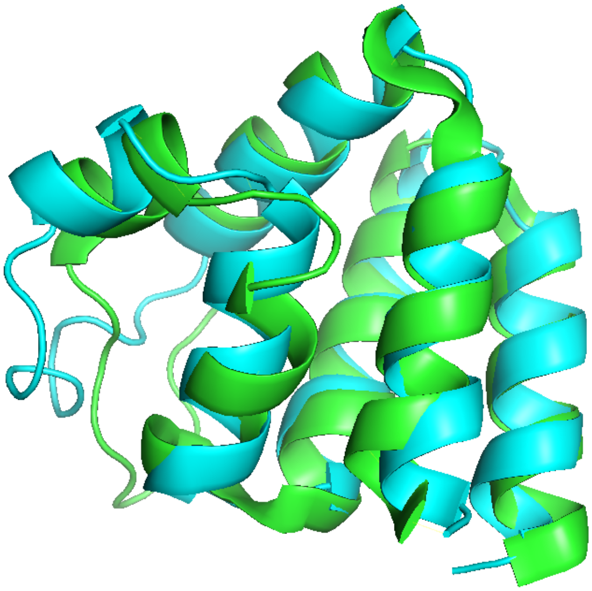


Figure 3: The best predicted conformation of 2MPC with the 6th state of the ground truth coordinates of 2MPC. RMSD = 2.60 Å

In response to this, RMSD values were recalculated using a method we developed, factoring in all states of the original coordinates. Furthermore, the best 10 RMSD scores, as well as the smallest RMSD and the matching state of the original conformation, were documented for each protein. These measurements were taken across different numbers of predicted conformations (10, 100, and 1000).

Table 4 presents the results of the best predicted protein conformations with the smallest mean RMSD compared to all states of the ground truth protein. It

pdb_entry	run_num	min_mean_RMSD	best_state	min_RMSD_structure	min_RMSD
2JYV	10	5.6687963244	6	2JYV_10_0007	5.3762436692
2JYV	100	2.889796624	4	2JYV_100_0093	2.5849254506
2JYV	1000	3.067017334	4	2JYV_1000_0059	2.5655332268
2MPC	10	4.7445736298	2	2MPC_10_0008	4.498485885
2MPC	100	4.1856320657	3	2MPC_100_0089	3.992239566
2MPC	1000	2.6064312099	6	2MPC_1000_0053	2.4028245376
2MS8	10	7.4903670296	12	2MS8_10_0002	7.0814274566
2MS8	100	4.3998273458	18	2MS8_100_0099	4.0022640553
2MS8	1000	3.7742616655	7	2MS8_1000_0813	3.2615478098
2M0D	10	5.0173995735	4	2M0D_10_0004	4.0478528583
2M0D	100	3.8767165118	11	2M0D_100_0069	3.3050146592
2M0D	1000	2.0770254063	18	2M0D_1000_0249	1.6249936226
6MWM	10	11.1530588	19	6MWM_10_0008	11.0508184034
6MWM	100	7.2796141565	15	6MWM_100_0011	7.0625705772
6MWM	1000	5.7380785105	7	6MWM_1000_0174	5.4174648295

Table 4: Best predicted conformations for every protein of Rosetta Ab Initio

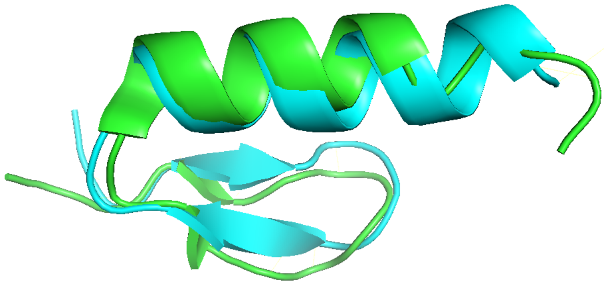


Figure 4: The best predicted conformation of 2M0D with the 18th state of the ground truth coordinates of 2M0D. RMSD = 2.07 Å

is evident that the most favorable RMSD values are predominantly observed in the 1000 predicted conformations. This observation aligns with the nature of the Monte Carlo method employed by Rosetta, as it introduces randomness into the quality of the outcomes. Consequently, a larger number of predicted conformations increases the probability of obtaining improved results. PyMOL was primarily utilized for the visualization of the results. Figures 2, 4, 3, 5 and 6 showcase the outcomes, demonstrating that regardless of the complexity of the protein, mostly favorable results were obtained.

In addition to analyzing the overall outcomes, we further examined the top 10 predicted conformations to assess the performance of different conformation numbers (10, 100 and 1000 conformations) and draw more conclusive insights regarding their efficiency. The y-axis represents the mean values of the 10 best predicted conformations. This analysis was conducted for all proteins, although Figures 7, 8, and 9 specifically illustrate the results for proteins 2JYV, 2M0D, and 2MS8, respectively. These proteins were chosen to represent different scenarios: 2JYV is a simplistic protein, 2M0D achieved the best RMSD, and 2MS8 is a more complex protein. It is evident from the figures that increasing the number of predicted conformations enhances the likelihood of obtaining more accurate predictions closer to the ground truth protein.

It is also interesting to observe that there is a significant difference in the mean RMSD between using 10 conformations and 100 conformations. Furthermore, there is a noticeable difference in the mean RMSD between 100 conformations and 1000 conformations, although the difference is not as pronounced as between 10 and 100 conformations. Typically, Rosetta Ab Initio is utilized on supercomputers, generating approximately tens of thousands of conformations to

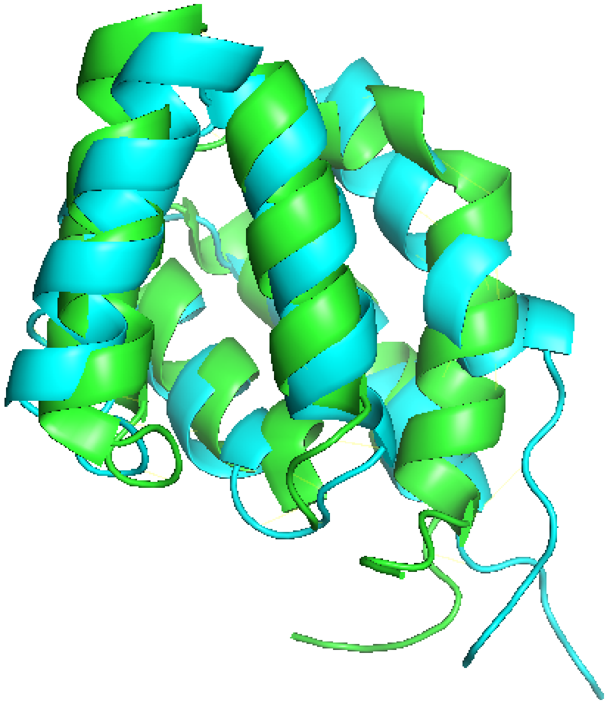


Figure 5: The best predicted conformation of 2MS8 with the 7th state of the ground truth coordinates of 2MS8. RMSD = 3.87 Å

ensure reliable outcomes. Unfortunately, we did not have access to a supercomputer for this study. However, based on our results, we believe that the final conformations obtained were sufficiently accurate for the purposes of our study.

4.2 Linear Algebra Algorithm

To prepare the input for the algorithm, a random state from the 1ADX and 1ANP ensembles was chosen (state 2 for both cases) and exported as new pdb files using PyMOL. Then, we ran our code to output their corresponding asymmetric Cayley-Menger matrices in Matlab format. Then both file were used as input in the Linear Algebra algorithm, which run with no errors, producing the new symmetric Cayley-Menger matrices.

Curiously, when checking the rank of the new ma-

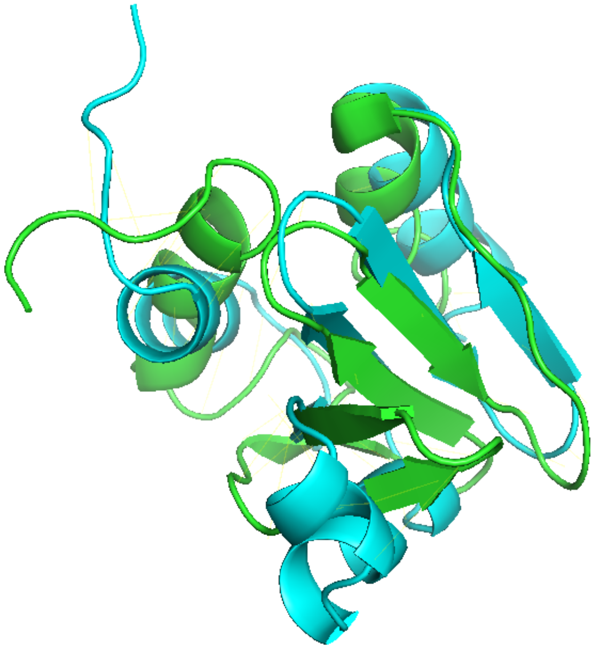


Figure 6: The best predicted conformation of 6MWM with the 7th state of the ground truth coordinates of 6MWM. RMSD = 5.74 Å

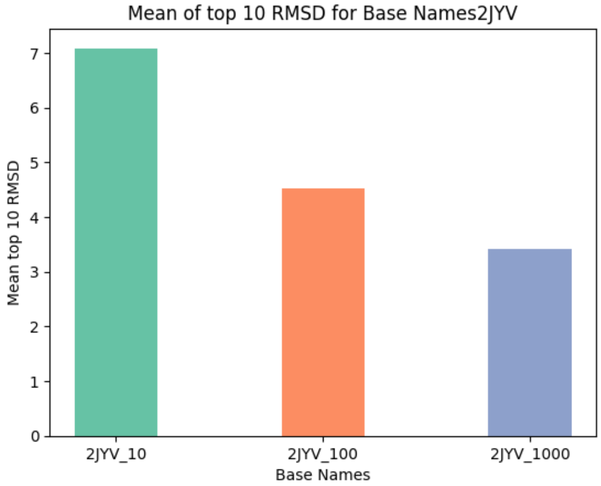


Figure 7: Barplot of mean RMSDs for 2JYV

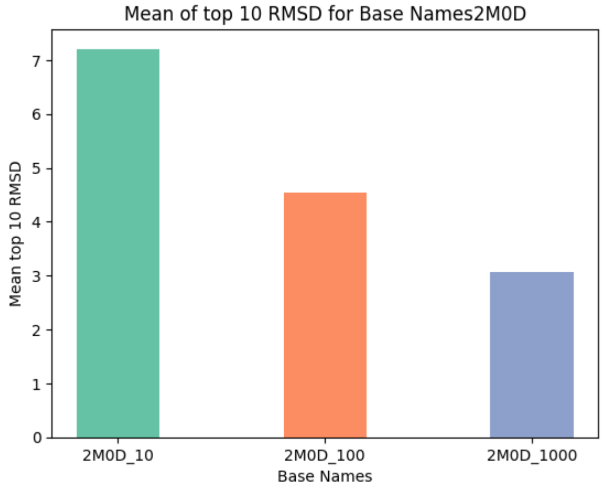


Figure 8: Barplot of mean RMSDs for 2M0D

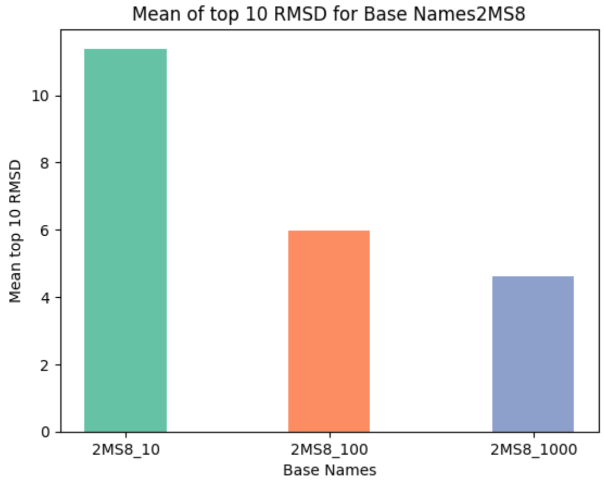


Figure 9: Barplot of mean RMSDs for 2MS8

trices in Matlab, both of them had a rank equal to their size, meaning that their rank was not reduced as expected. Even when we tested the algorithm with the `bounds7.m` file, an 8x8 asymmetric matrix that was included as an example in the directory of the Matlab algorithm with the rest of the files, its rank also failed to be reduced.

We proceeded to output the new Cayley-Menger matrices as csv files for both proteins from Matlab, and we run the rest of our code to revert them to distance matrices, extract their coordinates and calculate the RMSD between them and their corresponding state from the original pdb files. These results are shown in Table 5.

PDB ID	Chosen State	RMSD
1ADX	2	59.87611443959621
1ANP	2	41.37902340372047

Table 5: Linear Algebra Algorithm Results

Obviously, these results are not what we were looking for. To interpret them, we have to keep some things in mind. This Linear Algebra algorithm, which is designed to minimize the sixth singular value and thus reduce the rank of a matrix to 5, operates within certain constraints that can affect its outcome. These constraints include the structure and boundaries of the original matrix, the numerical precision of computations, and the stopping criteria for the algorithm. The perturbation process, guided by singular value decomposition, works within the limits set by the original matrix, and any restrictions imposed by these limits could prevent full rank reduction. The algorithm may also be influenced by computational and precision issues, such as rounding errors and the limitations of floating-point arithmetic. Additionally, the algorithm's stopping criteria, where it ceases operation when the sixth singular value is less than a specified tolerance, can also impact the final rank if the tolerance is not sufficiently small. Therefore, despite the algorithm's goal of rank reduction, in practice, it is possible for the final matrix to possess a rank larger than 5 due to these combined factors.

5 Discussion

Our investigation into protein structure prediction methods, namely the Rosetta Ab Initio[14] and a Linear Algebra Algorithm[10], revealed key insights and limitations regarding their application. Predicting protein structures is a cornerstone of modern biology, guiding drug discovery and disease understanding at the molecular level. Despite being alternative approaches to resource-intensive experimental methods like X-ray crystallography and cryo-electron microscopy, these computational methodologies present distinctive strengths and weaknesses.

The Rosetta Ab Initio algorithm, employed for predicting protein tertiary structures from amino acid sequences alone, demonstrated a correlation between the number of predicted structures and the accuracy of predictions. Our study noted that an increase in predicted structures resulted in smaller RMSD values, thereby implying better accuracy. Particularly effective for small to medium-sized proteins, Rosetta Ab Initio contributes significantly to our understanding of protein folding, interactions, and design. However, the algorithm was computationally intensive and its efficacy dropped with larger proteins. Ab initio methods, as exemplified by Rosetta Ab Initio, face the daunting task of dealing with an exponentially expanding search space for potential structures as protein length increases. The challenge is further exacerbated by predicting long-range interactions between distant amino acids, ultimately reducing accuracy for larger structures.

In contrast, the Linear Algebra Algorithm, which utilizes data from NMR spectroscopy to predict protein structures, proved inefficient in handling larger proteins due to its high computational complexity and extended runtimes. The results were also less satisfactory, with RMSD values significantly higher than anticipated. The algorithm's outcomes were subject to multiple constraints such as the structure and boundaries of the original matrix, numerical precision of computations, and the stopping criteria, which could impact its overall performance.

In conclusion, while both Rosetta Ab Initio and the Linear Algebra Algorithm provide crucial tools for predicting protein structures, they cannot be compared or combined due to their individual limitations. The choice of method would thus depend on the size of the protein under investigation and the computational resources available. Despite their respective constraints, the optimization of these methods or the development of new techniques, perhaps utilizing advancements in machine learning and computational science, holds the potential for considerably improved protein structure prediction, catalyzing breakthroughs in fields like drug discovery and disease understanding at the molecular level. The intriguing challenge of high accuracy prediction for larger proteins persists and calls for further research in this area.

6 Bibliography

[1] Ab-initio and ab-relax protocols. https://labnotes.readthedocs.io/en/latest/doc/ab-initio.html.

[2] Pymol documentation. https://pymol.org/dokuwiki/.

[3] Rosetta ab initio. https://www.rosettacommons.org/docs/latest/application_documentation/structure_prediction/abinitio.

[4] Rosetta ab initio flags. https://csrosetta.chemistry.ucsc.edu/node/1133, 2012.

[5] List of rosetta command line options. https://new.rosettacommons.org/docs/latest/full-options-listabrelax, 2022.

[6] Justin T. Catt Xuan Tan Robert G. Hammond Margaret A. Johnson Andrew J. Staup, Ivon U. De Silva. Structure of the sars-unique domain c from the bat coronavirus hku4. *Search on PubMed*, May 27, 2019.

[7] Benedetta A Sampoli Benitez, Michael J Hunter, David P Meininger, and Elizabeth A Komives. Structure of the fifth EGF-like domain of thrombomodulin: an EGF-like domain with a novel disulfide-bonding pattern 1 1edited by p. e. wright. *Journal of Molecular Biology*, 273(4):913–926, nov 1997.

[8] Bedard M. Bilodeau J. Lavigne P. Bernard, D. Nmr structure note: Structure of miz-1 zinc fingers 5 to 7. *TO BE PUBLISHED*.

[9] Anna Vinogradova Ping Wang Zhigang Chen Ping Xu Hugh P.J. Bennett Andrew Bateman Feng Ni Dmitri Tolkatchev, Suneil Malik. Structure dissection of human progranulin identifies well-folded granulin/epithelin modules with unique functional activities. *Search on PubMed*, 02 January 2009.

[10] Ioannis Z. Emiris and Theodoros G. Nikitopoulos. Molecular conformation search by distance matrix perturbations. *Journal of Mathematical Chemistry*, 37(3):233–253, apr 2005.

[11] Wayne J. Fairbrother, Robert S. McDowell, and Brian C. Cunningham. solution conformation of an atrial natriuretic peptide variant selective for the type a receptor. *Biochemistry*, 33(30):8897–8904, aug 1994.

[12] Sitao Wu Jooyoung Lee and Yang Zhang. Ab initio protein structure prediction. *Springer Science + Business Media B.V.*, 2009.

[13] Mumdooh Ahmed Lichun He, Benjamin Bardiaux. Structure determination of helical filaments by solid-state nmr spectroscopy. *Search on PubMed*, January 5, 2016.

[14] Mukundh Murthy. Ab initio protein folding. https://mukundh-murthy.medium.com/ab-initio-protein-folding-be27509d134a, 2020.

[15] Sarah J. Smith Qi-Rui Ong Stephanie L. Soh Katryn J. Justine M. Hill Parimala R. Vajjhala, Sebastian Kaiser. Identification of multifaceted binding modes for pyrin and asc pyrin domains gives insights into pyrin inflammasome assembly. *Search on PubMed*, August 2014.