EuroCC@Türkiye

This document is prepared by EuroCC@Türkiye for EuroCC2 under GA NO 101101903

Navigating Energy Surface of Functional Proteins

1. Problem Identification

In 2021, a very exciting development happened in the protein world. AlphaFold2 (AF2) program developed by DeepMind predicted, with unprecedented success, the three-dimensional conformation of proteins using only amino acid sequence information [1]. Although there were those who touted this as a solution to the "Protein Folding Problem", since the method is based on the examination of folded proteins of known structures with advanced machine learning techniques, the underlying physics still maintains its secrets. In this project, we are more concerned with what AF2 cannot do than what it can do.

Our group's research program is based-on predicting the functions of proteins whose three-dimensional structures are known using dynamical information obtained from trajectories and to foretell how the shifts in the environmental conditions will change these functions. Most folded proteins occupy multiple conformations that are close on the conformational space (CS) but are separated by high energy barriers. Computational and experimental methods determine only one or a few of these at best. Moreover, transitions between these occur on the timescales that are beyond the reach of most current computers. To further compicate matters, minor changes in environmental variables also affect the dynamics of transitions between these structures and their populations.

The aim of this project is to work with undergraduate students in the development of an efficient and integrated method that maps transitions between functional conformations of selected folded proteins to its energy surface. While studying protein dynamics students will explore the conformational space of proteins and we aspire to determine how changes in the environmental conditions modify this space. Our studies require extensive molecular dynamics simulations using enhanced sampling techniques. For this purpose, students have already learnt to run molecular dynamics simulations on open source software. They are expected to write their own analysis codes in Tcl and Python. The project require them to work with large amounts of data, consisting mainly of the coordinates of the protein systems.

Students have also successfully applied the metadynamics protocol to navigate between the minima of the protein calmodulin. In this project we will enhance the findings to determine not only the thermodynamics of the energy surface of this protein, but also the nature of the kinetic barriers using established protocol. Selecting the collective variable is the most crucial step in these types of problems. We will combine the infrequent metadynamics method [2] with the perturb-scan-pull approach developed in our group [3] and assess the success of the

methodology. We will also implement umbrella sampling as an alternative method to assess our findings.

Another protein of interest will be the protein TEM-1 β-lactamase which is the predominant source of resistance to penicillins in bacteria. The wild-type enzyme is an excellent penicillinase but has little activity against third generation cephalosporins, such as ceftazidime (CAZ) or cefotaxime (CTX). It was previously shown that this enzyme possesses cryptic allosteric sites; i.e. transient pockets in a folded protein that are invisible to conventional experiments but can alter enzymatic activity via allosteric communication with the active site. These sites offer promising opportunity for facilitating drug design by expanding the repertoire of available drug target positions. It is also known the dynamics of TEM-1 may be modified, e.g. by point mutations. However, a systematic mapping of its conformational surface under a range of conditions is yet to be carried out. In particular, we are interested in determining the so-called cryptic sites, that offer themselves to drug binding a small percentage of the time [4].

The third system of interest is the study of Green Fluorescent Protein (GFP). To investigate the functional and structural consequences of single residue mutations within the protein sequences, formerly generated scripts and the Mutation – Minimization (MuMi) scheme [5] previously developed in our group will be adopted to the study of GFP. With the integration of these outputs, the results will provide support for analyzing and understanding the relationship between protein structures, functions, and environmental factors, which will enhance the understanding of protein dynamics and provide significant information for areas such as structural biological research and therapeutic applications.

2. First Suggestion

The tentative project plan is as follows:

Timeline (months)	**Tasks - *Milestones***
1-5	Standard and accelerated MD simulations of TEM-1 β-lactamase – *determination of cryptic sites*
1-5	Metadynamics runs of calmodulin (CaM) – *conformational surface of CaM under four separate conditions*
1-5	MuMi of GFP – *fitness landscape for all point mutations of GFP*
4-6	Analysis of results; completion of missing runs, or supplementary runs for which need might arise during the analysis stages – *scripts developed for the analysis of the large data generated to be uploaded on our groups Github page (https://github.com/midstlab)*

3. Solution Stage – I

Zeynep Dilara Balkan has carried out accelarated MD (aMD) simulations on the TEM-1 β-lactamase system. She had already run local jobs for shorter simulations on computers at Sabanci University. She then chose 6 different sets of parameters to compare the outputs. She

optimized the parameters for the simulations she carried out on the HPC resources. She used local computers to analyze the outputs and to chart the conformational surface of the TEM-1 β-lactamase system.

Durmuş Erdem Kertmen performed the MuMi scheme on all residue positions of wild type GFP (PDB:1EMA). Mutated coordinate files (pdbs) were generated on local computers with a python script (mumi1.py) which uses the Prody package [6]. Structure files (psf) for MD simuations were created with the 'autopsf' tcl script which was run on VMD software on the HPC. After that, configuration files containing the parameters for the minimization were genereated with another python script (mumi2.py). This script also writes out a file (config_joblist) containing the names of all configuration files as inputs so that the all minimization runs can be submitted as a single job. To run the minimization, 64 CPUs and 1 GPU was allocated on a single node. Each minimization had a benchmark time of around 0.003 sec/step, yielding a total of 30 sec for 10000 steps of minimization. Whole run was finished in around 35.9 hours. Finally, another python script (mumi3.py) was used to parse the last frame of each trajectory. For that, Prody package was installed on HPC by creating an Apptainer container as described here. Scripts used in this work are shared as a notebook in our Github page. Last frames are downloaded to a local computer and to be analyzed in terms of solvent accessible surface area and hydrogen bonding around the chromophore environment.

Dilara Coban and Sila Horozoglu studied the free energy surface of calmodulin (CaM) using the metadynamics methodology. They prepared four sets of CaM simulations, each exploring the calcium bound and calcium free forms at low versus physiological ionic strengths. Using the HPC resources, they started out with 100 ns simulations and then prolonged the simulations until convergence on the surfaces were obtained. They determined that 500 ns of well-tempered metadynamics simulations are sufficient to chart the conformational surface of this molecule. They also determined the regions of similarities and differences between the surfaces and made physics-based explanations for their observations.

Each student has learned to submit jobs on and retrieve outputs from the HPC. They have analyzed the outputs on lcoal computers.

4. Results and Achievements

All these enhanced simulations require extensive computational resources. The undergraduate students in this project were trained to effectively use HPC resources, benchmark their jobs, and optimize computer resources to solve the problem at hand. They also learned to write analysis scripts to handle the large amount of output produced on the resources.

Some of the findings in this study were presented as a poster at the 23rd European Conference on Computational Biology (https://eccb2024.fi/):

B. Tayhan, S. Horozoglu, D. Coban, A.R. Atilgan, C. Atilgan, "Protean Nature of Calmodulin: Mapping Conformational Dynamics Across Environmental Shifts"

5. HPC Benefits

The students had already learned to prepare the simulations and to submit jobs on our local cluster at Sabanci University. Through this use case, they scaled up their runs and learned to prepare their systems to best fit the allocated resources on an HPC. The students who worked in this project were from three different majors: (i) Computer Science and Engineering, (ii) Molecular Biology, Genetics and Bioengineering, (iii) Materials Science and Nano Engineering. This was a strength in that they shared experiences and benefited from each other's knowledge. Their common goal was to do graduate studies in computational sciences at top universities around the World. The HPC experience they gained through this project was an important asset for them to reach these goals; at the same time the project contributed to the development of human resources in Türkiye, in particular those doing graduate studies in our group who helped mentor the undergraduates. We hope to diffuse the knowledge we gain in the various dissemination activities of the EUROCC2 project as well as by sharing the codes developed in open-source environments.

The students who were a part of this project:

Sıla Horozoğlu sila.horozoglu@sabanciuniv.edu
Dilara Çoban dilara.coban@sabanciuniv.edu
Durmuş Erdem Kertmen ekertmen@sabanciuniv.edu
Zeynep Dilara Balkan zeynepbalkan@sabanciuniv.edu

The graduate students who co-advised the students:

Büşra Tayhan busra.tayhan@sabanciuniv.edu
Melike Berksöz melike.berksoz@sabanciuniv.edu

6. Challenges

The one big challenge of working with undergraduate students is that they do not have a grasp of the resources that are offered to them. They associate the term 'high performance computing' with infinite resources. They therefore do not plan their jobs according to the limitations of the system. It is important to lead them into benchmarking their runs and to optimize what is offered to them. They also need to understand they might have to redefine their goals in connection with the available resources.

7. References

[1] J. Jumper et al., “Highly Accurate Protein Structure Prediction with AlphaFold,” Nature, 596, 583-589 (2021). https://www.nature.com/articles/s41586-021-03819-2

[2] D. Ray and M. Parrinello, “Kinetics from Metadynamics: Principles, Applications, and Outlook,” J. Chem. Theory Comput., J. Chem. Theory Comput., 19, 5649-5670 (2023). https://doi.org/10.1021/acs.jctc.3c00660

[3] F. Jalalypour, O. Sensoy, C. Atilgan, "Perturb-Scan-Pull: A Novel Method Facilitating Conformational Transitions in Proteins," J. Chem. Theory Comput., 16, 3825-3841(2020). https://pubs.acs.org/doi/10.1021/acs.jctc.9b01222

[4] A. Kuzmanic et al. “Investigating cryptic binding sites by molecular dynamics simulations.” Acc Chem Res 2020, 53:654–661. https://pubs.acs.org/doi/10.1021/acs.accounts.9b00613

[5] G. Ozbaykal, A.R. Atilgan, C. Atilgan, “In Silico Mutational Studies of Hsp70 Disclose Sites with Distinct Functional Attributes,” Proteins: Structure, Function, Bioinformatics, 83, 2077-2090 (2015). http://dx.doi.org/10.1002/prot.24925

[6] Bakan A, Meireles LM, Bahar I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011** 27(11):1575-1577.

Summary:

In 2021, AlphaFold2 revolutionized protein prediction, but it does not solve all protein folding mysteries. This project aims to explore what AlphaFold2 cannot do, focusing on predicting protein functions and how environmental changes affect them. Students will develop methods to map transitions between protein conformations, using molecular dynamics simulations and enhanced sampling techniques. They'll analyze proteins like calmodulin and TEM-1 β-lactamase, exploring cryptic allosteric sites and dynamics under various conditions. They'll also study Green Fluorescent Protein mutations. This project requires extensive computational resources and will train students in high-performance computing and analysis scripting.

Özet:

2021'de AlphaFold2 protein tahmininde devrim yarattı, ancak bu devrim tüm protein katlama gizemlerini çözmüyor. Bu proje, protein fonksiyonlarını tahmin etmeye ve çevresel değişikliklerin onları nasıl etkilediğine odaklanarak AlphaFold2'nin yapamadıklarını keşfetmeyi amaçlıyor. Öğrenciler, moleküler dinamik simülasyonları ve gelişmiş örnekleme teknikleri kullanarak protein konformasyonları arasındaki geçişleri haritalamak için yöntemler geliştirecekler. Kalmodulin ve TEM-1 β-laktamaz gibi proteinleri analiz ederek çeşitli koşullar altında kriptik allosterik bölgeleri ve dinamikleri keşfedecekler. Ayrıca Yeşil Floresan Protein mutasyonlarını da inceleyecekler. Bu proje kapsamlı hesaplama kaynakları gerektirmektedir; öğrencileri yüksek performanslı hesaplama ve analiz komut dosyası yazma konusunda eğitecektir.