Master 2 Physics Photonics Nanotechnology 2022-2023

Functional modes and thermal B-factor predictions for multigenic structural analysis predicted from Alpha Fold

Nicolas PETIOT

May 1, 2023

Supervisors: Dr. Adrien NICOLAÏ and Pr. Patrick SENET

NANOSCIENCES Department - Physics Applied to Proteins

Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS
Université de Bourgogne Franche-Comté / Faculté des Sciences et Techniques Mirande
9, Av. Savary - B.P. 47 870 21078. DIJON CEDEX - France

Abstract

Proteins are complex biomolecules that are critical to the functioning of living organisms. They are made up of sequences of amino-acids that fold into specific three-dimentional structures, which is highly related to their function. The process of protein folding is considered one of the most challenging problems in the field of biology and biochemistry, as it involves a delicate interplay of chemical and physical forces that determine the final shape of the protein. AlphaFold is a groundbreaking tool developped by DeepMind that uses artificial intelligence algorithms to pedict 3D structure of proteins based on their amino-acid sequence. This tool has the potential to revolutionize the study of enzymes as it provides fast and accurate way to predict the structure of molecules that have catalytic properties. The present work aim at using AlphaFold to study the catalytic properties of Glutathione Transferase (GST), especially from class δ and ε of *drosophilia melanogaster*, whith the ultimate goal of being able to design brand new sequences for enzymes with improved catalytic efficiancy.
In addition to using AlphaFold for the generation of 3D structures, molecular dynamics simulations can also be performed based on these structures. It allows to predict some of their potential behaviour and study various factors that may influence the function of the protein, such as thermal fluctuations or protein-ligand interactions. Simulationg these processes gives a deeper understanding of how proteins function and allows to identify areas for further investigation and improvements.

Contents

1 Introduction

1.1 Glutathione Transferase

Glutathione Transferase (GST) is a superfamily of enzymes that are generally homodimeric structures and multigenic in numerous organisms. They are involved in detoxification process as well as in chemoperception in mammals and insects[1]. Their main function is to catalyze the conjugation of reduced glutathione (GSH) to xenobiotic electrophilic centers[2]. In their catalytic cycle, the GSH usually binds in a specific set of amino-acids called G-site and the hydrophobic xenobiotic in the so-called H-site. Interactions between insects and plant's chemicals lead to a major driving force in herbivorous insect evolution, hence this encourages the study of insect GSTs to understand how spontaneous mutations modify the stability, selectivity and the catalytic efficiencies of this enzyme superfamily. In the present work, we have selected 25 sequences of GST from *drosophilia melanogaster*, 11 from class δ and 14 from class ε. This will constitute a working base for all the incomming analysis.

1.2 AlphaFold & X-ray diffraction experiment

X-ray diffraction is a powerful experimental technique that have been used extensively to determine the three-dimentional structures of proteins. In this technique, a crystal of the protein is bombarded with X-rays, and the resulting diffraction pattern is used to determine the position of atoms within the protein. Over the years, X-ray diffraction experiments have played a pivotal role in determining the structures of tens of thousand of proteins, which are deposited in the Protein Data Bank (PDB). However, this process can be time-consuming and technically challenging. Moreover, compared to the vast number of known protein sequences, the ensemble of solved structure is insignificant. In 2021, DeepMind used machine learning approaches with the AlphaFold[3] programm. It uses computational models to predict the 3D structures of proteins based on it's sequence with a high accuracy. In the field of de novo design of enzymes, ALphaFold has the potential to revolutionize the way we consider the design process, allowing to predict 3D structures that have not yet been experimentally characterized. In addition to the initial AlphaFold program, DeepMind developped several other tools that have further expanded the capacities of protein structure predictions. One such tool is AlphaFold-multimer[4], which allows predictions for the structure of protein complexe such as homodimers. An other one is AlphaFill[5], which predict the positions of ligands, small molecules that bind to protein such as Glutathione. All together, these tools represent a major step forward in the field of protein study and will be at the root of the present work.

1.3 Goals

2 Materials and Methods

2.1 Multiple Sequence Alignment

Multiple sequence alignment (MSA) is a fundamental technique in bioinformatics used to compare and analyze the similarities and differences between multiple biological sequences. These sequences can be DNA, RNA, or protein sequences and can come from various species or different regions of the same genome. By aligning these sequences, it is possible to identify conserved regions that are important for function or evolution, as well as unique features that differentiate the sequences. In this study, we focus on a set of 25 GSTs sequences related to each other through darwinian evolution. Each sequence has a different length, making the alignment process especially useful. Through this analysis, we aim to identify regions of conservation and divergence between the sequences (see exemple bellow). Our first task was to use MSA to predict the position of the dimer interface and binding site of the set of sequences. The stability of the dimer structure is dependent on the interactions at this interface. Therefore, understanding the location and conservation of the dimer interface can provide insights into the stability of the dimer structure and the mechanisms of the biological activity. On the other hand, the binding site is the specific location on the enzyme where a substrate or ligand binds and interacts. The catalytic efficiency of the enzyme is related to the binding site because it determines the specificity and strength of the substrate-enzyme interaction. Therefore, understanding the location and conservation of the binding site is essential for elucidating the function and mechanism of action of the enzyme.

Each cell in the MSA matrix corresponds to a particular amino acid at a particular position in a particular sequence. We refer to the position in the MSA matrix as the MSA index. The MSA index allows us to compare the amino acid residues at each position across all the sequences in the alignment. We focused on highlighting residues that are known to be part of the dimer interface or binding site. By doing so, we can compare these residues across all sequences and identify any conserved or variable regions.

```
MSA sample :
seq1 ...gqk-eq-qlypk...
seq2 ...aks--d-elypk...
seq3 ...ans--d-elypr...
(conserved region)
```

2.2 Anisotropic Network Model

As seen in the introduction, the AlphaFold program allows for the prediction of protein structures with remarkable accuracy. These structures have the potential to be used in a wide range of applications, including molecular dynamics simulations. In this work, we aim to use AlphaFold-predicted structures as input for the Anisotropic Network Model (ANM) to study protein dynamics. Specifically, we will compare the results obtained from the ANM simulations using AlphaFold structures with those obtained using experimentally determined structures. More precisely, we will use the ANM to predict the thermal B-factors of the studied structures[6], which are important indicators of protein flexibility and stability. The insights gained from this study will contribute to the ongoing efforts to develop computational tools for protein structure analysis and facilitate a deeper understanding of protein

dynamics and function. To achieve these goals, it is necessary to provide a detailed description of the ANM and its underlying mathematical principles. The ANM is a widely used method for studying the collective motions and dynamics of proteins based on their structure. It models the protein as a network of connected nodes and springs (representing covalent and non-covalent interactions between them).
Let $\vec{r_i}$ being the position of the node i and M_i it's mass. In the ANM, each node is assumed to be at the bottom of an harmonic potential, since interactions are modeled by connections between nodes, the force matrix is obtained by computing the mass-weighted Hessian matrix

$$\hat{H}_{ij} = -\frac{\Gamma_{ij}\gamma}{\sqrt{M_i M_j}} \frac{\vec{R_{ij}}\vec{R_{ij}^T}}{R_{ij}^2} \tag{2.1}$$

where γ is the spring constant used to model interactions between nodes, $\vec{R_{ij}} = \vec{r_j} - \vec{r_i}$, and Γ is the contact matrix. In the case were $i = j$, the force matrix is computed so that the self interacting term is the response to all the applied forces.

$$\hat{H}_{ii} = -\sum_{j \neq i} H_{ij} \tag{2.2}$$

Γ_{ij} is equal to 1 if we consider a connection between the nodes i and j and 0 else. The computation of Γ is as follows, given a cutoff radius R_c, two nodes i and j are connected by a spring if $|\vec{R_{ij}}| < R_c$. The normal modes and eigenfrequencies are given by the diagonalization of the mass-weighted Hessian matrix.

$$\hat{H}\vec{e}_k = \tilde{\omega}_k^2 \vec{e}_k \tag{2.3}$$

It is important to make some remarks at this point, first in the equation (2.1) $\hat{H}_{ij}$ is actually a three by three matrix. It means that the ovearall $\hat{H}$ will be a square matrix of dimention $d = 3N$ with N the number of considered nodes. Since the diagonalization algorythms have a complexity of $O(d^3)$, it means that the computation time will highly depends on the choosen set of nodes. Second, the masses of the nodes are expressed in g.mol^{-1} to avoid numerical errors in the diagonalization. Expressing γ in J.m^{-2}, it means that the eigenvalues $\tilde{\omega}_k^2$ are expressed in kmol.s^{-2}. The eigenfrequency expressed in Hz is then given by $\omega_k = \sqrt{\mathcal{N}_a \times \tilde{\omega}_k^2 \times 10^3}$, where $\mathcal{N}_a$ is the constant of Avogadro. Thermal B-factors of the node i can be deduced from the normal modes using the following

$$\beta_i = \frac{8\pi^2}{3} \frac{k_B T}{M_i} \sum_k \frac{|\vec{e}_{k,i}|^2}{\tilde{\omega}_k^2} \tag{2.4}$$

where $\vec{e}_{k,i}$ contains the elements of $\vec{e}_k$ related to the node i. β_i is expressed in m^2 but will always be converted in ^{2} because of conventions and usual order of magnitude.
Ever since the beginning, we were talking in a very abstract way about "nodes". In this work, we did consider several ensembles of nodes, namely the amino-acid's center of mass (COM), the heavy atom's position (ATM) and the atom's position (ATM+H). This allowed us to analyse structures with a various number of node. When available, we compared the values computed with the various ANMs with X-ray based measurement in term of pearson correlation.

$$\mathcal{R} = \frac{\sum_i (\beta_i - <\beta>_i)(B_i - <B>_i)}{\sqrt{\sum_i (\beta_i - <\beta>_i)^2} \sqrt{\sum_j (B_j - <B>_j)^2}} \tag{2.5}$$

The values of $\mathcal{R}$ are between -1 and 1 and gives the linear correlation between predicted and measured B-factors. A coefficient of 1 is associated to a perfect correlation wherease a coefficient of 0 means that there is basically no links between prediction and experiment. Eventually, a negative value of $\mathcal{R}$ means that there is a correlation with opposite sign and would be interpreted as a result even worst than no correlation. Note that we need to consider the same ensemble of points for β_i and B_i.

Bibliography

(1) Gonis, E.; Fraichard, S.; Chertemps, T.; Hecker, A.; Schwartz, M.; Canon, F.; Neiers, F. *Insects* **2022**, *13*, 612.

(2) Nicolaï, A.; Petiot, N.; Grassein, P.; Delarue, P.; Neiers, F.; Senet, P. *Applied Sciences* **2022**, *12*, 8196.

(3) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. *Nature* **2021**, *596*, 583–589.

(4) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J., et al. *BioRxiv* **2021**, 2021–10.

(5) Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A. *Nature Methods* **2022**, 1–9.

(6) Atilgan, A. R.; Durell, S.; Jernigan, R. L.; Demirel, M. C.; Keskin, O; Bahar, I. *Biophysical journal* **2001**, *80*, 505–515.