Numerical simulations of proteins

Ramirez Palacios, Carlos

DOI:
10.33612/diss.196789826

Chapter 10

Summary

Computer simulations allow the study of biomolecular systems in time and length scales otherwise unreachable by experimental techniques. An important aspect to creating computational models is that they must make testable predictions. Validation ensures that the numerical models represent reality and that we have a good understanding of the mechanisms wherefrom observable properties arise. In this thesis, I use three numerical methods to model biomolecular systems (proteins): molecular dynamics, molecular docking, and deep learning. The first two methods treat the system as a set of particles that interact according to preset rules (force field or energy function). Deep learning methods are data-driven and the system is treated depending on the architecture and type of representation used. While molecular dynamics and docking are well-stablished methods with a plethora of existing literature and expertise, deep learning is an emerging field with the potential to revolutionize the way we simulate biomolecular systems. In short, this work is about three computational methods to model proteins applied to catalysis. Details on experimental work for each chapter can be found in the respective publication.

In **Chapter 2**, a computational protocol for predicting the enantioselectivity of *Vf*-TA is developed and validated on a benchmark dataset of compounds. We show that modelling one intermediate from the transamination reaction is enough to make accurate predictions about the global transamination reaction. A combination of docking and molecular dynamics enables better sampling of the conformational space, and yields an accurate but computationally inexpensive algorithm for modelling the *Vf*-TA selectivity. Additionally, it is shown that there is a correlation between the Rosetta interface energy and the catalytic activity of *Vf*-TA, which could be exploited in enzyme engineering efforts. This idea is further explored in **Chapters 3** and **4**.

In **Chapter 3**, the Rosetta interface energy was used as the objective function to guide the redesign of a thermostable *Pj*-TA variant. The *Pj*-TA mutants were modelled computationally, and experimental collaborators carried out the laboratory verification of the designed mutants. The results confirmed that the proposed method for modelling ω-TAs can aid protein engineering efforts; not only was the hit rate high (68 out of 70 variants exhibited catalytic activity) but also the measured yields were in accordance with the expected values. Further rationalization of the structure-activity relationship was presented using the docked complexes. Finally, ns-scale molecular

dynamics simulations were performed on the docked complexes to measure the water displacement upon substrate binding, and a weak correlation across mutants was observed. The purpose of the work presented in this chapter is twofold: to validate a computationally-inexpensive methodology for redesigning the substrate scope of *Pj*-TA, and to show the potential of the thermostable *Pj*-TA variants in the synthesis of chiral amines.

Chapter 4 is a small chapter showing more datasets where the enzymatic activity of *Vf*-TA and *Pj*-TA could be predicted by using the Rosetta Interface Energy. Rational analysis of the docked structures was performed, and we show that even though *Vf*-TA and *Pj*-TA are similar in the identity of the residues found in the binding site, the conformations that those residues go through to accommodate a new substrate are different and that is what makes their reaction profiles different. The work is further evidence that *Pj*-TA has the potential of a broader substrate spectrum.

In **Chapter 5**, it is shown that the, at the time, newly-developed Martini 3 force field can be used in unbiased binding of small organic molecules into the binding site of proteins. In the chapter, the binding of ligands to two enzymes are described (*Vf*-TA and Src-kinase), and the full manuscript contains seven showcase systems. The application is exciting because it opens up the possibility of studying binding paths, which are computationally too expensive for atomistic simulations. As mentioned in Chapter 3, there starts to be an interest in modelling not only the Michaelis complex but also the process that leads to its formation and the role of the solvent in said process. Additionally, it indicates that the Martini 3 force field is accurate enough to model interactions between small organic molecules and proteins. Ensuring the accuracy of Martini is important because it brings confidence that simulations of much larger systems (e.g., an entire cell) represent reality.

Chapter 6 is a small chapter showing the molecular modelling of yet another PLP-dependent enzyme, a diaminopimelate decarboxylase (DAPDCs) from *Thermotoga maritima*, to rationalize the observed activity profiles. Some of the redesigned DAPDCs catalyzed the conversion of 2-aminopimelic acid to 6-aminocaproic acid, a building block in the synthesis of nylon-6. The challenges in modelling decarboxylases were that barely any publications on the structure-activity relationship were available and that the fact that the reaction mechanism is not completely elucidated. We hypothesized that the C_α-CO_2^- cleavage would be the limiting step in the reaction, and thus modelled the external aldimine intermediate of the decarboxylation. The shift in catalytic activities of the E315X mutants was rationalized on the basis of the occurrence of binding poses found in molecular dynamics trajectories. Some correlation between frequency of reactive binding poses and enzymatic activity is shown, but there is room for improvement. In **Chapter 9**, it

is shown that the analysis of molecular dynamics trajectories can be done with very little manual intervention and could aid in the definition of binding poses of interest.

Chapter 7 is about the molecular modelling of an *N*-glycosyltransferase from *Actinobacillus pleuropneumoniae* (ApNGT). This chapter is strongly based on results from experimental methods. The experimental collaborators had established that ApNGT catalyzes the glycosylation of adhesin fragments in a semiprocessive manner, and molecular modelling was used to try to describe the mechanism from which processivity arises. The challenges in modelling ApNGT were that no crystal structure in complex with glucose or the peptide (adhesin fragment) is known, and that the reaction mechanism is not well understood. Rosetta was used to model the protein-peptide or ligand-peptide complexes. We hypothesized that the processivity arises from the shallow substrate binding groove in ApNGT, and a sliding mechanism is proposed.

In **Chapter 8**, it is shown that it is possible to train a spectral graph convolutional neural network to predict the binding energies of combinatorial libraries of *Vf*-TA mutants. The neural network was trained on a small dataset of mutants with binding energies obtained via traditional molecular modelling methods. In a few hours, the neural network learned the intricate synergic interplays between mutation sites and became accurate in evaluating unseen variants. The usefulness of this approach is that the trained neural network is several orders of magnitude faster than the traditional molecular modelling method at screening new variants, bringing the time needed to scan one variant from ~20 min to <1 millisecond. The short evaluation time potentially enables the screening of the entire search space. Furthermore, it is shown that the accuracy of the predictions can be further increased by injection of feature embeddings obtained from a pretrained language model. To the best of my knowledge, this is the first time such an approach is reported in combinatorial libraries.

In **Chapter 9**, the molecular dynamics trajectories generated in **Chapter 2** are revisited to show the potential of using neural networks to analyze trajectories. Similar to what was presented in **Chapter 2**, we have a set of simulation trajectories and would like to classify them into “reactive” or “non-reactive” depending on the enantiomer occupying the binding site of the enzyme. The trained models achieved high accuracy in telling the two classes apart. This chapter is more demonstrative than applicative. I demonstrate that using a neural network to identify geometric criteria of interest could reduce the manual intervention needed to pull off strategies such as the ones from **Chapters 2** or **6**.

As the reader can tell, the common theme across chapters was the modelling of proteins by numerical methods. The PhD work was not limited to any one particular subject or research question, but rather to explore the possibilities of computational

modelling. The majority of the publications are the result of collaborations with experimentalists, which greatly serves as validation that the computational methods represent reality. Generally, the purpose of experimentalists was to tailor the enzyme to accept some desired substrate, and my purpose was to design algorithms to predict and model the outcome of the enzymatic reaction. Later in the thesis, the reader is introduced to machine learning methods to model or aid the modelling of proteins.

Numerical simulations of molecular systems are becoming steadily better at reproducing reality. This has been made possible by both the increase in computing power in recent decades, the development of more efficient algorithms, and a better understanding of the mechanisms that govern molecular interactions. Research groups investigating molecular systems are increasingly more interested in also reproducing their work via computational simulations. Nevertheless, it is important to continue innovating, and I would argue that the recent advances in the machine learning field will greatly facilitate it, and may even be the push needed to reach the next step in the simulation of reality.