Studying Changes in Protein Pathway Accessibility with Motion Planning

Abigail Ren, Diane Uwacu, Shawna Thomas, Nancy M. Amato***

[1] Department of Computer Science,
Vassar College, Poughkeepsie, New York, USA,
abigailren@vassar.edu.
[2] Parasol Lab, Department of Computer Science and Engineering,
Texas A&M University, College Station, TX, USA,
{duwacu, sthomas, amato}@tamu.edu.

Abstract. The study of protein-ligand binding is an important area of research in drug design. When a ligand, a small drug molecule, binds to a protein at an active binding site, the protein will undergo structural changes to make it more difficult for the ligand to escape. In this paper, we use motion planning to model protein-ligand binding in order to study the changes in accessibility that occur in the protein structure after binding with a ligand. We experimented on two pairs of proteins and developed an initial idea for a method to score accessibility of access tunnels and proteins. Overall, we found that protein accessibility decreased from the unbound state to the bound state. Accessibility was affected negatively by fewer valid access tunnels and higher energy levels in the bound protein state compared to those in the unbound state.

1 Introduction

The study of protein-ligand interaction is important to research on drug molecules. A ligand, which can be considered a small drug molecule, interacts with a protein at a binding site. When bound to a protein, the ligand can trigger or inhibit a reaction, making it important that the ligand only interacts with the right protein. This is a concern in the study of drug design, as many drugs cause side effects by interacting with the wrong protein. By studying protein-ligand binding, we hope to be able to better understand how the ligand binds to the protein and what factors will influence a ligand to have higher affinity with one protein over another.

While binding sites can occur at the surface of the protein, we focus on proteins with buried binding sites. These are found deep within the protein,

* This research supported in part by NSF awards CNS-0551685, CCF 0702765, CCF-0833199, CCF-1439145, CCF-1423111, CCF-0830753, IIS-0916053, IIS-0917266, EFRI-1240483, RI-1217991, by NIH NCI R25 CA090301-11,and by DOE awards DE-AC02-06CH11357, DE-NA0002376, B575363.
** This work was performed at the Parasol Lab during Summer 2019 and supported in part by the CRA-W Distributed REU (DREU) project.

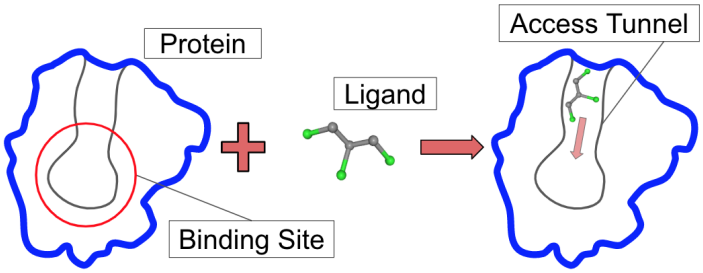


Fig. 1: Protein-Ligand Binding

and the ligand will need to pass through access tunnels in order to reach the binding site, as shown in Figure 1. While much research has been done on the active binding site, research on the tunnel pathways to the binding site is still relatively new. We know that protein structures will change after protein-ligand binding, often to prevent the ligand from escaping easily once it has bound to the protein. We hypothesize that the accessibility of the active binding site decreases in the ligand-bound protein, potentially through factors like fewer valid pathways to travel through or higher energy levels that inhibit ligand mobility. Because of this, we focused on the study of protein tunnels to gather more data.

One way to study these tunnels is through the application of motion planning to computational biology. Motion planning has been used in this area of research to model how a drug molecule might interact with a protein to access its binding site. This is done by using the protein structure as a complex 3D environment and the drug molecule, or ligand, as the robot. For this purpose, this project uses motion planning to model the access pathways in a given protein and study the changes that occurred from the unbound protein to the ligand-bound protein. We will refer to these two states as the unbound and bound state respectively.

In this work, we will use existing sampling-based motion planning based approaches to model protein tunnels for a given protein and analyze our results with a method that will score the accessibility of the tunnels found and the protein itself. We implemented an initial idea for an accessibility scoring function that takes into account the number of valid pathways found inside a protein and the energy levels along each pathway.

2 Related Work

In this section, we will explain the biological concepts motivating this paper and the motion planning features that we utilized in our code.

2.1 Protein-Ligand Binding

Recently, more research is being done based on the Keyhole-Lock-Key Model, which takes into account how the structure of protein tunnels or "keyholes" can

affect protein-ligand binding activity. The protein structure can change depending on whether it is bound to a ligand or not. In bound proteins, formerly valid pathways may close up to prevent the ligand from the binding site, but also to prevent other ligands from entering as well.

Previous research focused heavily on the active binding site without consideration for how the tunnels to the binding site might also affect the ligand's ability to reach the binding site. Tunnel geometry [5] can contribute to ligand binding not only by restricting the size and shapes of the ligands that are allowed through to the binding site, but also by shaping the interactions and barriers within tunnels that may influence how a ligand is able to reach the binding site.

2.2 Guided Planning

Due to the complexity of protein structures, it can be difficult for sampling-based motion planning algorithms to navigate the free space of the environment and find valid paths through the many narrow passages within the protein structure. Dynamic Region-Biased Rapidly-Exploring Random Tree (DRRRT) [3] works well in this environment because it uses a skeleton that represents the topology of the free space of a protein to guide itself through the narrow areas of the inside of the protein. The algorithm chooses regions guided by the skeleton and samples nodes with the given region. Once enough valid configurations are generated, the algorithm is then able to continue sampling a new region along the skeleton, allowing the roadmap to have a more representative sample of valid robot configurations within the environment.

To further increase optimization in finding paths through the protein, we can allow connectivity using Rapidly-Exploring Random Graphs (DRRRG) [4]. This method generalizes the tree generated in DRRRT into a graph structure and creates a more comprehensive sample of valid node configurations of the environment. Connectivity is increased in a graph structure and results in the finding of more optimal paths than the construction of a tree.

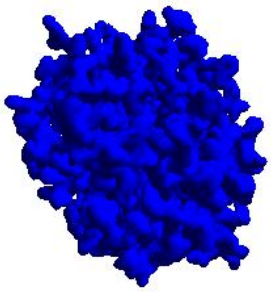
3 Overall Approach

In this section, we will describe the methods that we used to model accessible ligand pathways in protein and to gather data on the protein tunnels.

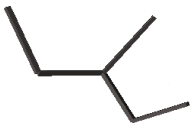
3.1 Modeling

We generated geometric models of the protein and ligand for our experiment. We use the protein structure for our experiment environment and treat the ligand structure as the body of our movable robot. Although proteins are dynamic structures that will change as a ligand moves through it, these are minor changes in structure that are negligible. Therefore, for the purposes of our experiment, we utilize rigid protein structures in the environment.

Figure 6 shows an example of the protein and ligand models that were used in our experiments.



(a) 4FWB



(b) TCP

Fig. 3: Geometric models of protein (a) and a ligand (b)

3.2 Planning

In the planning stage, as shown in Figure 4, we first generated a Mean Curvature Skeleton (MCS) [8] of the protein structure to represent the free space of the protein. We used the generated skeleton to guide the motion planning sampler to build the planning roadmap.

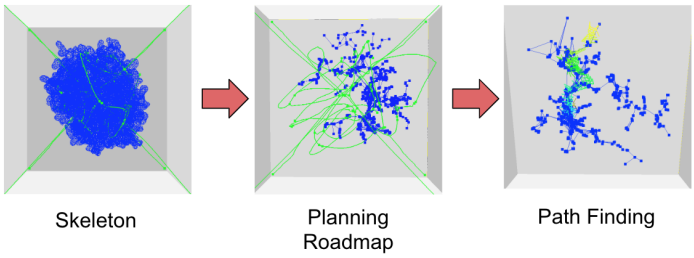


Fig. 4: Planning Phase

Our motion planning method is based off the sampling-based methods defined in Section 2.2. We use a Mean Curvature skeleton guided RRG [4] to sample valid configurations of the ligand robot within the protein environments. This is used to construct a roadmap of the valid robot configurations, which is used in the path finding process. By generating start and goal nodes, we can query through the roadmap to generate possible paths that the ligand robot can take from the surface of the protein to the active binding site. Once any region has been explored thoroughly, we end the planning method to obtain the valid paths generated by each seed.

3.3 Analysis

Once all the pathways are generated, we analyze them further to determine which seeds in the experiment discovered the same tunnels. We check how similar the discovered tunnels are to each other and cluster the ones that pass a certain similarity threshold in order to retain distinct pathways.

(a) Paths Found by all Seeds

(b) Distinct Tunnels after Path Similarity Comparison

Fig. 6: After performing analysis on paths found by the motion planning algorithm, we are able to distinguish which pathways can be clustered together and treated as the same pathway.

Finally, we implemented a scoring function to score the accessibility of tunnels and proteins.

The tunnel score takes into consideration the volume and energy feasibility within each tunnel. This is demonstrated by metrics like the number of seeds that discovered a particular tunnel and the energy levels within the tunnel. The following is the format for the tunnel score:

$$\text{Num. Seeds} \cdot w_0 + \text{Min Energy} \cdot w_1 - \text{Max Energy} \cdot w_2 \tag{1}$$

The more seeds that discover a tunnel, the greater positive factor it has on the tunnel scoring method. This is an indicator of tunnel volume and represents how likely the pathway was found by the motion planning method by the number of seeds that discovered the tunnel. A tunnel can be seen as easier to plan through when more seeds discover valid paths through it. Energy levels also play an important role in accessibility. Lower energy levels help a ligand traverse through a protein tunnel, whereas high energy levels can act as a barrier that inhibits a ligand's ability to pass through a tunnel. For these reasons, smaller minimum energies of a tunnel have a positive influence on the accessibility score, whereas high energy levels detract from the score, especially if they near the maximum energy threshold. The scoring function takes the average minimum and maximum energies found across seeds for a particular tunnel.

Similar to the tunnel scoring function, the protein score is formatted as follows:

$$\text{Num. Tunnels} \cdot w_0 + \text{Avg. Tunnel Score} \cdot w_1 + \text{Min Energy} \cdot w_2 - \text{Max Energy} \cdot w_3 \tag{2}$$

In addition to energy levels, the protein score also considers the total number of distinct tunnels that were found for the protein. If the number of tunnels discovered pass the expected tunnel count threshold, the weight of the number of tunnels found is maximized. The method also considers the average tunnel score of the protein. This provides a general idea of how accessible the protein tunnels within the protein might be.

Using these two functions, we assess the accessibility of the tunnels and proteins and produce a score on a scale from zero to one in order to provide a statistic to judge protein pathway accessibility.

4 Experiments

4.1 Experimental Setup

We used PDB files from the RCSB Protein Data Bank [2] to model protein structures for our experiment. We also used the software UCSF Chimera [7] to generate geometric models of the protein and the ligand. Using these structures, we built our experiment environment, using the protein structure as a 3D obstacle in the configuration space and the ligand as the movable robot within the environment.

When running the experiment, we tested with six seeds and took the average of the data generated by all the seeds. We gathered information about the roadmap construction run time, which provides information about the difficulty of planning the roadmap. We also gathered data on the number of tunnels that were found as well as the minimum and maximum energy levels in each tunnel and averaged those values respectively to use in our scoring function.

Parameters that we took into consideration included setting the threshold for energy at 1000 in order to ignore found tunnels whose energy levels were very high. We also have a threshold for the number of expected pathways per protein. We gathered this from literature and research done about each protein that identified access tunnels with the protein and indicated how many were found. We used this to influence the scoring function so that the weight of the tunnel volume wouldn't be overcompensating when the found number of protein pathways from our experiments exceeded the number of tunnels that we expected to find. This kept the tunnel volume weight within a fixed interval for the scoring function.

In order to study the cost and quality of the roadmap by the motion planning algorithm, we extracted information about the run time of constructing the roadmap. We also looked at the number of tunnels that were found by the motion planning algorithm. From this, we also gathered data about the energy

levels in the protein tunnels. Using these metrics we applied the scoring function to produce an accessibility score for the tunnels and proteins.

We gathered statistics about the quality and cost of roadmap construction for each protein and scored the tunnels that were found to produce an overall protein score. The tunnels and proteins were scored based on accessibility. We took into consideration the number of tunnels that were found as well as the minimum and maximum energy levels measured within the protein as explained in equation 2.

4.2 Results

In Figure 7, the run time for constructing the roadmap increased for HCV NS5B. This indicates that the motion planning algorithm expended more time trying to find valid robot configurations in the free space of the environment, suggesting that the bound state of the protein was a more difficult environment to build a roadmap in. DhaA31 did not follow this trend, as the run time decreased, suggesting that the bound protein was an easier environment to plan through.

Figure 8 displays the difference in the number of tunnels found between the unbound and bound state of the protein. We can see that the number of tunnels decreases for HCV NS5B, which again contributes to the accessibility score because there are fewer pathways being generated for a ligand to traverse through. This makes it more difficult for a ligand to access a binding site because the ligand now has to search harder in order to find a valid pathway in the bound state. Meanwhile, the number of tunnels for DhaA31 stayed the same.

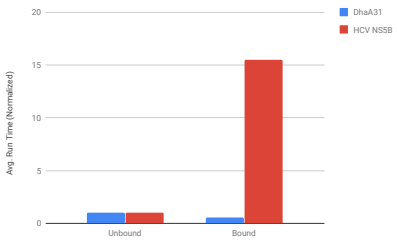


Fig. 7: Avg. Run Time

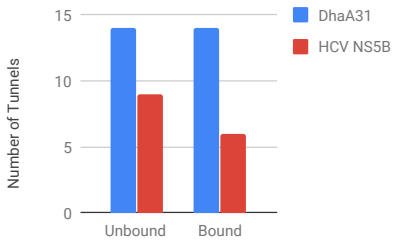


Fig. 8: Num. of Tunnels

Protein	Score	
	Unbound	Bound
DhaA31	0.50	0.48
HCV NS5B	0.54	0.26

Table 1: Protein Accessibility Score

Using the above statistics, we applied them to the scoring function to get a final protein score for each protein. As shown in Table 1, the protein accessibility score decreased from the unbound protein state to the bound state for both sets of proteins. This indicates that it was more difficult for the ligand to access the active binding site within the protein when it was in the bound state. This is obvious for HCV NS5B, as the protein underwent structural changes that indicated decreased accessibility to the active binding site. However, for DhaA31, there were no obvious structural changes from the data we gather that suggest the bound protein had lower accessibility. The decrease in score instead comes from the differences in energy levels that we gathered from our data, as the energy levels within the bound protein of DhaA31 were higher in general, making it less conducive for the ligand to move through the protein.

5 Conclusion

Overall, we observed that pathway accessibility decreased from the unbound state to the bound state. From our data, we can infer that this is affected both by fewer accessible pathways for the ligand to find the binding site, and higher energy levels that inhibit a ligand's ability to progress from a particular pathway. This provides us with some insight into how protein structures change when interacting with a ligand and will help guide our future work. We plan to expand the current data set with more unbound/bound protein pairs. We will also work on implementing a more sophisticated scoring system and improve upon the current scoring method. In order to compare our results against existing methods of protein accessibility, we intend to use programs like DINC [1] and Probis [6]. These programs analyze the binding site in order to evaluate protein accessibility, and we hope to show that even when the binding site does not change much between the unbound and bound protein, our method is able to detect and score the changes in accessibility due to our consideration of access tunnels in the protein.

6 Acknowledgments

I would like to thank my mentors Diane Uwacu, Dr. Amato, and everyone at Parasol Labs for helping me this summer. This research was sponsored by CRA-W DREU and the University of Illinois at Urbana-Champaign.

References

1. Antunes, D.A., Moll, M., Devaurs, D., Jackson, K.R., Lizée, G., Kavraki, L.E.: Dinc 2.0: a new protein-peptide docking webserver using an incremental approach. Cancer Research 77, e55–57 (2017)
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., Bourne, P.: The protein data bank. Nucleic Acids Research 28(1), 235–242 (2000)

3. Denny, J., Sandstrom, R., Bregger, A., Amato, N.M.: Dynamic region-biased exploring random trees. In: Proc. Int. Workshop on Algorithmic Foundations of Robotics (WAFR). San Francisco, CA (December 2016)
4. Kala, R.: Rapidly exploring random graphs: motion planning of multiple mobile robots. Advanced Robotics 27(14), 1113–1122 (2013), https://doi.org/10.1080/01691864.2013.805472
5. Kingsley, L.J., Lill, M.A.: Substrate tunnels in enzymes: Structure-function relationships and computational methodology. Proteins 83(4), 599–611 (2015)
6. Konc, J., Miller, B.T., Štular, T., Lešnik, S., Woodcock, H.L., Brooks, B.R., Janežič, D.: ProBiS-CHARMMing: Web Interface for Prediction and Optimization of Ligands in Protein Binding Sites. Journal of Chemical Information and Modeling 55(11), 2308–2314 (2015), https://doi.org/10.1021/acs.jcim.5b00534, pMID: 26509288
7. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E.: Ucsf chimera: A visualization system for exploratory research and analysis. Journal of Computational Chemistry 25(13), 1605–16012 (2004)
8. Tagliasacchi, A., Alhashim, I., Olson, M., Zhang, H.: Mean curvature skeletons. Computer Graphics Forum 31(5), 1735–1744 (2012)