Predicting Ligand Binding Sites with UOBPRM and Machine Learning

Anthony Enem, Benjamin Porter, Diane Uwacu, Shawna Thomas, Nancy Amato

***Abstract*— Many approaches to predicting ligand binding sites on protein surfaces suffer from a reliance on unreliable and inaccurate evaluations of potential binding sites. In this work, we propose an approach to ligand binding that takes advantage of the motion planning paradigm of randomized sampling to uniformly generate samples of the ligand at various binding site candidates near the protein surface. We then compute metrics that describe the favorability of each configuration's location as a potential binding site. We also propose future work to apply these metrics to a machine learning algorithm to take advantage of each metric's strengths and weaknesses and reliably predict the locations of binding sites for ligands on proteins.**

I. Introduction

Ligand binding is the process by which a ligand (drug) binds to a specific pocket on a protein, where its atoms can create a stable reaction with the atoms of the protein, called a binding site. This process is useful in analyzing the efficiency of drug molecules. Ligands must be both shaped in a way that will make interactions with the protein easy and must react in an energetically favorable way.

In this work, we use motion planning sampling to create samples of the ligand around the protein. The locations of these samples are initially considered to be possible binding sites for the ligand. We then compute metrics that further narrow the possibilities of the locations of binding pockets.

Our research focuses on computing metrics that are used when given a set of input (ligand and protein) to predict ligand binding sites on the protein body. We also suggest an application for these metrics to a machine learning approach to more consistently predict possible binding sites based on the strengths and weaknesses of specific metrics.

II. Related Work

A. Motion Planning

Motion planning is a problem that has many applications to our world, from computer animations to robotic medical procedures and protein folding and ligand binding simulations. Its basic premise is to find a valid path from a start to a goal configuration, given an environment and descriptions of moveable objects (robots) and obstacles. The robot attempts to reach the goal configuration while avoiding collisions with the obstacles. Although this problem seems deceptively simple, it gets increasingly difficult as we introduce more restraints and degrees of freedom to the robots. Except for robots with very few degrees of freedom, the problem is computationally hard [5]. In modeling protein folding, the robot often has many degrees of freedom, making computations particularly difficult.

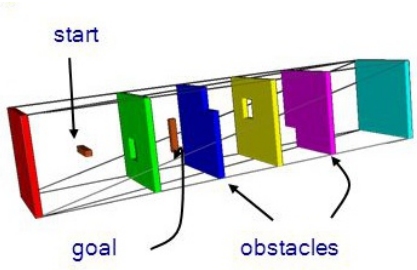

Fig. 1. 2D Maze: The robot must traverse from the lower left corner (in red) to the upper right corner (in blue) of the environment.

We make use of a randomized sampling algorithm (UOBPRM) to generate our ligand samples around the protein. We do not use the usual next step in motion planning, the probabilistic roadmap (PRM), which is constructed by connecting the configurations using a local planner [4]. Our project just makes use of the sampling mechanism of PRMs. However, some future work could be done with PRMs in the form of metrics that take into account whether a configuration generated with our method is actually reachable from the exterior of the protein. In this project, we model the ligand as a linkage and the protein as a rigid obstacle in the environment in the motion planning environment.

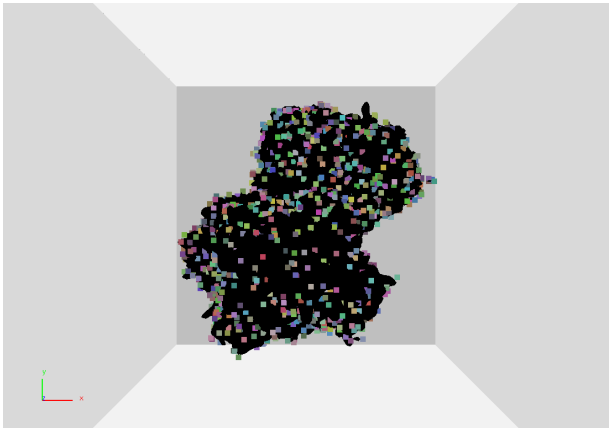

Fig. 2. UOBPRM: Generation of 2,500 samples shown in point mode around the 4RRW protein

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.

The work of Enem, Porter performed at the Parasol Lab during Summer 2016 and supported in part by the CRA-W Distributed REU (DREU) project.

1) Uniform Obstacle-Based PRM (UOBPRM): In generating samples around an obstacle, the Uniform Obstacle-Based PRM (UOBPRM) guarantees a uniform distribution of samples near C-obstacle surfaces [3]. UOBPRM has been demonstrated to have a better node distribution around $C-obst$ than the Obstacle-Based PRM (OBPRM) [1], and also better proximity of the samples to the obstacle than the uniform random sampling method. In order to identify binding pockets on a protein body, our generated ligand samples have to be both in close proximity to the protein body and uniformly distributed around the protein so that all possible areas of the protein surface are considered.

B. Machine Learning

Tom Mitchell of Carnegie Mellon University defines machine learning as follows: A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

We hope to apply the metrics generated by our strategy to a machine learning algorithm, particularly a neural nework where the metrics are supplied as features. The training phase consists of running the algorithm with ligand/protein sets for which the true binding sites are known in order to take advantage of the strengths of the metrics in different situations. Next, the algorithm would be tested on ligand/protein sets for which it does not know the true binding site location to measure its accuracy.

The two main types of supervised machine learning approaches are regression systems, which return a confidence measure on a spectrum of possible responses, and classification systems, which return a simple yes or no prediction. We propose a regression system to return a measure of likelihood for each sample to be located within a binding pocket.

III. Overal Approach

In this work, we first generate samples with the randomized sampling method UOBPRM, then calculate metrics for each sample using a motion planning strategy. Finally, we propose future work to apply these metrics to a machine learning algorithm to greatly improve its planning capability by taking into account the different metrics' strengths in different circumstances.

A. Studied Metrics

In order to overcome the difficulties of successfully predicting ligand binding sites using traditional methods, we calculate a variety of metrics which describe characteristics of each sample, such as how close to or buried it is in the protein or the energy of the configuration according to van der Waals interactions. We also present future work in integrating the strengths and weaknesses of each metric in a machine learning approach.

Algorithm 1 AnalyzeStrategy to calculate metrics for each sample

Input. map: file containing all samples generated by UOBPRM
Output. Metrics for use with machine learning algorithm
  for each $sample \in map$ **do**
    $metrics \leftarrow sample_{ID}$
    $metrics \leftarrow$ `Distance`$(sample_{COM}, Protein_{COM})$
    $metrics \leftarrow$ `Energy`$(sample, ProteinAtomCoordinates)$
    $metrics \leftarrow$ `ConvexHullScore`$(sample)$
    $metrics \leftarrow$ `PQPSolidPenetrationDepth`$(sample)$
  end for
  return $metrics$

1) Distance to the Center of Mass: The first and simplest metric that we calculate is the distance between the center of mass of the protein and the center of mass of the ligand sample. The distance to the center os mass is usually lower for samples which are buried in pockets on the protein surface (Figure 3). This calculation is a normalization of the resultant of subtracting two vectors which represent the coordinates in 3D space of the protain and ligand center of masses. Although binding pockets are generally located relatively close to the center of mass of the protein, this metric alone is not sufficient to determine the presence of a site.

Fig. 3. Distance to COM: Showing distances of genrated samples to protein's center of mass (marked X).

2) Energy: We calculate the energy of each configuration using the equation proposed by Levitt in [6]. This equation accounts for the average energy between each pair r_{ij} of the ligand center of mass and a nitrogen, calcium, or carbon atom from the protein. In our calculations, we ignore the value of the hydrophobic interactions.

TABLE I
VAN DER WAALS CONSTANTS FOR DIFFERENT ATOMS

Atom	A	B
N	395280	2556
Ca	3075695	953
C	1200965	425

Fig. 4. The Levitt energy equation

$$U = \sum_{atompairi,j} A/r_{ij}^{12} - B/r_{ij}^{12} + E_{hydrophobic}$$

Algorithm 2 `Energy`

Input. $ligand$: the configuration to calculate energy for, $atoms$: a vector containing the x,y,z coordinates of all N, Ca, and C atoms of the protein

Output. The average energy of the configuration

$energy \leftarrow 0$
for each $atom \in atoms$ **do**
  $r_{ij} \leftarrow$ `Distance`$(atom, ligand_{COM})$
  set values of van der Waals constants A and B
  $energy = energy + A/r_{ij}^{12} - B/r_{ij}^{6} + E_{hydrophobic}$
end for
return `Average`$(energy)$

3) Convex Hull Clearance and Penetration Scores: Another metric we computed was which samples are buried in cavities on the protein body, and how deep the samples are buried in those cavities. To implement this, we created a convex hull around the protein body and determined samples that were in collision with the convex hull.

Samples that are not covered by the convex hull were assigned a score of 0, samples partially covered are assigned a score of 1, and samples fully covered are assigned a score of 2 (Figure 5). The greater the score assigned to a sample, the more likely the location of the sample is to be a binding pocket. The distance of each sample to the body of the convex hull is computed as a depth value. For samples fully covered by the convex hull, this value is negative. For samples partially covered, the value is 0. For samples outside the convex hull, the depth value is computed as a positive value.

IV. Experiments

We ran two sets of experiments to verify our choice of UOBPRM as the optimal sampling method for our approach and to gain insight into how our well our metrics agree with one another, respectively.

A. Sampling method comparison

To verify our choice of sampling method, we generated 1,000 nodes using uniform random sampling (UniformRandomFree), obstacle-based sampling (OBPRM), and uniform obstacle-based sampling (UOBPRM). When comparing the output, UOBPRM clearly produces a uniform distribution around the obstacle surface, which is most desirable for locating binding pockets.

B. Measure of confidence in metrics

To get a measure of confidence for our metrics, we ran tests on the 3W6H protein and the Zn^{2+} ligand using our implemented metrics.

Samples buried in pockets of proteins are more likely to have lower distances to the protein's center of mass than most other samples. Although the distance metric is meant to take this into consideration, it is only helpful for bulky shaped proteins. For linear shaped proteins, this metric would not be as useful.

TABLE II
Average distance to nearest neighbor node for 1,000 samples

Sampler	Distance
UniformRandomFree	11.5491
OBPRM	2.1958
UOBPRM	2.2102

For this experiment, our energy function may have suffered from inaccuracies because we did not consider hydrophobic interactions, substituting a zero for their value. The energy function was therefore dependent only on the distance between the ligand and the atoms of the protein.

The binding site scores from the convex hull exhibit little variation among the samples generated because the convex hull metric only considers collisions betweeen the samples and the convex hull of the protein and because there are only three possible scores, causing the same scores to be reported for large number of samples. This metric only takes into account the geometric properties of the protein when assigning scores to samples, so it is also not very useful on its own in identifying a binding site on a protein for a specific ligand.

V. Results

We ran tests which generated 1,000 samples around the 3W6H protein using three different samplers (Uniform Random, Obstacle-Based PRM, and Uniform Obstacle-Based PRM). The purpose of these tests was to determine the best sampling method for our project. The results can be seen in table II. While uniform random free sampling produces a distribution with a large distance between neighbor nodes, OBPRM and UOBPRM produce much smaller and closer average distances. Because UOBPRM's value is smaller, it suggests that it exhibits slightly more clumping in its node distribution, making UOBPRM the optimal choice.

The uniform random sampler generated samples uniformly in the environment, but far away from the protein body (Figure 6). The obstacle-based sampler generated samples close to the protein. Although the close proximity to the protein is useful for this project, OBPRM oftens generates samples clustered on parts of the protein, leaving some parts uncovered (Figure 7). The uniform obstacle-based sampler generated samples which are both close to the protein and uniformly distributed around the protein body (Figure 8). These results demonstrate that UOBPRM is the best choice for our sampling method.

Our distance metric relies on the geometry of the protein body. It proves to be more useful for bulk shaped proteins. For other irregular shaped protein bodies (e.g linear), the distance metric would not be as useful. This is due to the fact that the cavities on the surface of the proteins would be smaller for linear shaped proteins. Also, the distance from the center

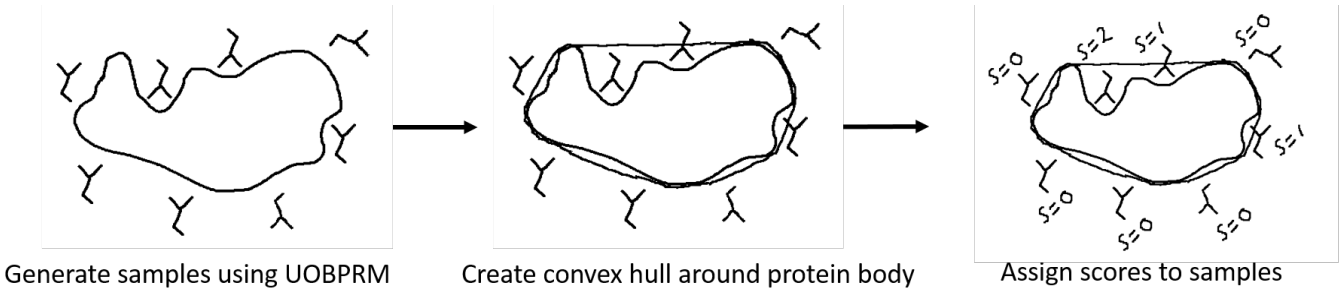

Fig. 5. Convex Hull Generation: Showing the process of generating the convex hull and assigning scores to samples.

TABLE III
BEST AND WORST CASE SAMPLES FOR ALL METRICS WITH 3W6H PROTEIN AND Zn^{2+} LIGAND

Metrics	Best/Worst COM Distance		Best/Worst Energy		Best/Worst Depth	
Best/Worst	Best	Worst	Best	Worst	Best	Worst
COM Distance	3.2377	41.3307	34.888	28.033	15.751	34.645
Energy	6.42e-2	-2.85e-4	-3.99e-8	2.62e7	3.03e-4	-4.28e-8
Depth Score	1	1	2	1	2	0

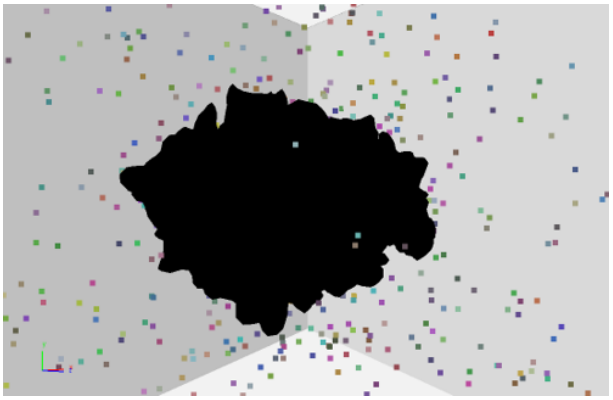

Fig. 6. Uniform Random Sampler node generation in point mode: 1000 samples are randomly generated in the environment and far from the protein body.

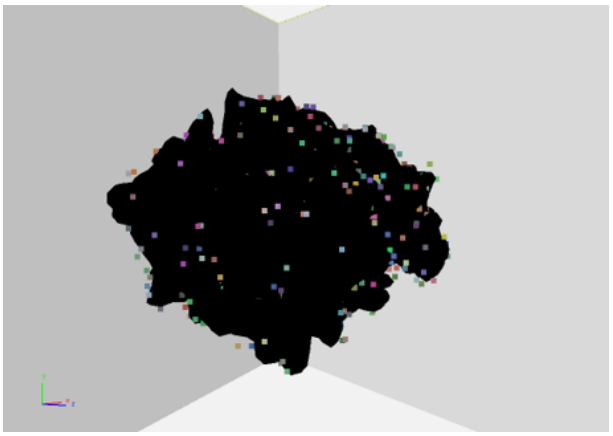

Fig. 7. Obstacle-Based Sampler node generation in point mode: 1000 samples are generated close to the obstacle but clustered on one side, leaving some parts of the protein uncovered.

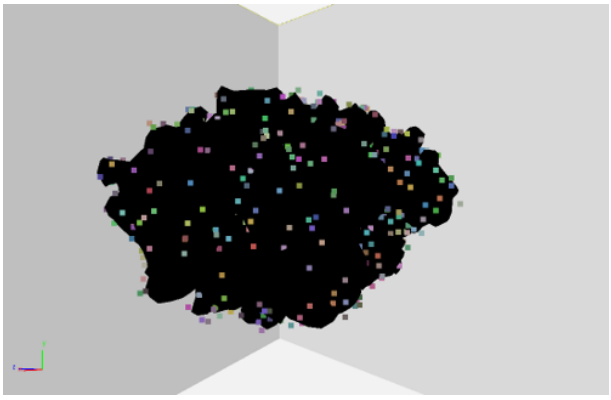

Fig. 8. Uniform Obstacle-Based Sampler node generation in point mode: 1000 samples uniformly generated around the protein surface.

of mass would be larger due to the length of longer linear proteins.

Since we do not take the hydrophobic interactions into consideration for this project, our energy values are solely dependent on the distance from the ligands to the atoms on the protein. From our energy function (Figure 4), we can see that the energy value gets smaller generally for larger distances between ligands and atoms of the protein. Because of this, the samples with lower energies tend to occur far away from the center of mass of the protein. Therefore, the distance metric and energy metric often disagree with one another.

VI. CONCLUSION

In this work, we propose a motion planning approach to evaluate candidate binding sites on a protein's surface using randomized sampling. We established that UOBPRM is the optimal choice of sampling method to achieve this due to its uniform sampling close to the protein surface. We also calculated metrics for each sample that quantify characteristics

of an ideal binding site. We determined that the metrics often do not agree perfectly with one another and sometimes disagree greatly. Therefore, we propose the implementation of a regression-style neural network with the metrics as features in order to take advantage of each metric to train the method and predict binding sites on proteins for which the real binding sites are not known.

VII. Acknowledgments

We would like to express our gratitude toward our faculty mentor, Dr. Nancy Amato, our postdoctoral mentor, Dr. Shawna Thomas, and our graduate student mentor, Diane Uwacu.

We also thank the faculty, graduate students, summer interns, and all members of the Parasol Lab we worked with during this project. We also thank the DREU program for assisting in the funding of this project, and our respective institutions for preparing us for this experience.

References

[1] N. M. Amato, O. B. Bayazit, L. K. Dale, C. Jones, and D. Vallejo. OBPRM: an obstacle-based PRM for 3d workspaces. In *Proceedings of the third Workshop on the Algorithmic Foundations of Robotics*, pages 155–168, Natick, MA, USA, 1998. A. K. Peters, Ltd. (WAFR '98).

[2] N. M. Amato, O. B. Bayazit, L. K. Dale, C. V. Jones, and D. Vallejo. Choosing good distance metrics and local planners for probabilistic roadmap methods. *IEEE Trans. Robot. Automat.*, 16(4):442–447, August 2000.

[3] H. Y. (Cindy), S. L. Thomas, D. Eppstein, and N. M. Amato. UOBPRM: A uniformly distributed obstacle-based PRM. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012, Vilamoura, Algarve, Portugal, October 7-12, 2012*, pages 2655–2662, 2012.

[4] L. E. Kavraki, P. Švestka, J. C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.

[5] J. C. Latombe. Motion planning: A journey of robots, molecules, digital actors, and other artifacts. *Int. Journal of Robotics Research*, 18(11):1119–1128, 1999.

[6] M. Levitt. Protein folding by restrained energy minimization and molecular dynamics. *J. Mol. Biol.*, 170:723–764, 1983.