Using Motion Planning to Study Ligand Binding

Ogenna Esimai, Hsin-Yi (Cindy) Yeh, Shawna Thomas, Nancy M. Amato
Parasol Lab, Department of Computer Science and Engineering, Texas A&M University, College Station, TX, USA
Email: {oesimai, hyeh, sthomas, amato}@cse.tamu.edu

Abstract—**In the drug discovery process, pharmaceutical companies screen many candidates for the most promising. Drug screening is costly, but by carrying out a part of it computationally or virtually, the cost can be reduced. An effective drug molecule acts as a ligand that binds to the active site of a protein to form a protein-ligand complex. The binding configurations of the protein and ligand may be predicted by molecular docking. The predicted configurations may be used to determine the binding affinity of the complex. Ligand binding and molecular docking are computationally intensive problems in computational biology. We present an approach that applies a motion planning technique to these problems. We approximate ligands to robots and proteins to obstacles using Uniform Obstacle-Based Probabilistic Roadmap (UOBPRM), a novel planning algorithm that uniformly distributes robot configurations around obstacle surfaces. We develop an algorithm to test and rank protein-ligand complexes based on an approximation of binding affinity. We present five complexes with experimentally determined binding affinities published in the literature. We find that our simulated ranking of these complexes matches the ranking from the published binding affinities. Thus, UOBPRM shows promise as a potential technique with which to rank protein-ligand complexes based on their binding affinity properties. This information may be useful as a cost-saving measure to pharmaceutical companies in the area of computational or virtual drug screening.**

I. Introduction

In pharmaceuticals, the screening process rigorously tests potential drugs to select the most promising candidates. The drug discovery process is costly [5, 11, 8, 9] and carrying out part of drug screening computationally or virtually may cut cost. The strength with which a drug binds its target is the binding affinity. The drug, called a ligand forms bonds with the protein yielding a complex and causing a desired effect. The process in which the ligand attaches to the protein is called ligand binding or docking. The preferred configurations of the ligand and the protein in their bound state may be predicted in a modeling process known as molecular docking [7, 24].

Molecular docking experiments are accurate but expensive and labor-intensive [22]. Therefore, many computational molecular docking mechanisms have been designed to reduce the cost by treating both protein and ligand molecules as rigid bodies such as DOCK[14], AutoDock[19], GOLD[6], and FlexX[18]. While useful, these programs may require specialized expertise to use and they are not very accurate. In the past, work in robotics has been applied to solving computational biology problems such as protein folding [1, 23] and ligand binding [3, 20]. Motion planning helps to find a trajectory for a robot from a start to a goal. By modeling molecules as robots, motion planning can be perfectly fitted to the study of molecule motions. Previously, a motion planning approach called Obstacle-Based Probabilistic Roadmap (OBPRM) [2] was used to generate potential ligand samples around the protein in a random fashion [3]. In that work, the ligand is assigned some flexibility by modeling it as a linkage robot. Furthermore, a motion planning algorithm and human user input are incorporated to help predicting ligand binding sites. The results show that user input largely augments the effective ability of the automated OBPRM motion planner to predict the true ligand binding site.

A limitation of OBPRM is that it is unable to guarantee the distribution of the generated robot samples. Uniform distribution is important since we can use the least amount of samples to represent the whole space. Recently, a new approach, UOBPRM (Uniform Obstacle-Based Probabilistic Roadmap) [25], was proposed to uniformly distribute robot samples around obstacle surfaces.

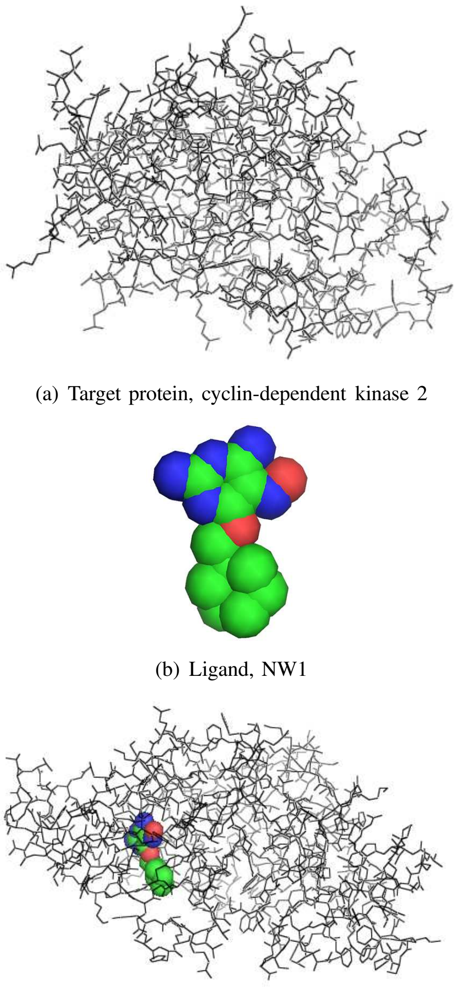
In this paper, we model the ligand as a linkage robot and the protein as an obstacle similar to [3] and use UOBPRM to generate ligand samples that are uniformly distributed around the protein surfaces. We approximate binding affinities for five protein-ligand complexes using the distance between the protein and the ligand and we rank these complexes based on the approximation of their affinities. Our ranking matches the ranking obtained from the experimentally-determined binding affinities published in the literature. Our results show that UOBPRM is useful as a potential technique for ranking protein-ligand complexes based on an approximation of binding affinity and this ranking can be useful for computational drug screening in the future.

II. Related Work

A. Ligand Binding

A ligand is a small molecule, for example, a drug. Ligand binding is the process in which a ligand (Figure 1(b)) comes in close proximity to its target, usually a protein (Figure 1(a)), and navigates to the part of the protein where it forms bonds with the protein resulting in a protein-ligand complex (Figure 1(c)).

Ligand binding can sometimes cause a subsequent change in conformation of the protein producing a specific effect, e.g., causing a cell membrane channel in the body to open and allow the passage of ions. The protein is usually many orders of magnitude larger than the ligand in size. The specific area in the protein to which the ligand binds is called the binding pocket or binding site or especially for proteins that are enzymes, the active site. The strength of the binding between the protein and ligand is called binding affinity. Binding



(a) Target protein, cyclin-dependent kinase 2

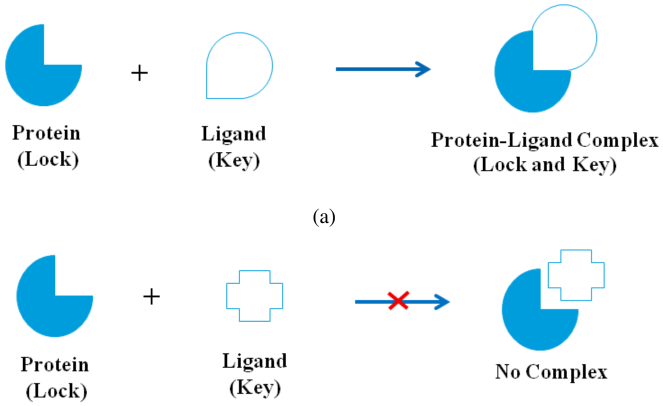
(b) Ligand, NW1

(c) Protein-ligand complex, 1E1X

Fig. 1. A ligand, NW1, binds with its target protein, cyclin-dependent kinase 2 to form a protein-ligand complex, 1E1X.

affinity is important because it is related to the stability of the resulting protein-ligand complex. The higher the binding affinity, the more stable the protein-ligand complex that is formed. Different proteins bind different ligands with different affinities. Multiple factors affect affinity, for example, a higher bond energy between the protein and the ligand, a greater ability of the ligand to reach and remain in the active site, and a longer time spent by the ligand in the active site, all contribute to higher binding affinity.

Ligand binding can be described using the lock and key model. The protein is the lock and the ligand the key in this model. Just as not all keys will fit a particular lock, not all ligands are able to bind to the protein. One example of this lock and key model is shown in Figure 2. Figure 2(a) shows that the ligand and the protein are able to form a protein-ligand complex when their shapes match. Figure 2(b) is an example of when the protein and the ligand do not have complementary shapes, a protein-ligand complex is not formed and no following mechanism occurs.



(a)

(b)

Fig. 2. Lock and key model. (a) A protein and a ligand with compatible geometry which successfully form a complex. (b) Here, the protein and the ligand do not have compatible geometry. As a result, no complex is formed.

B. Motion Planning

A robot is a movable object which can be represented by n parameters (degrees of freedom). Each parameter represents one object component, such as object position or object orientation. All possible robot placements or configurations form an n-dimensional space, C-space. Each robot can be represented as a point (x_1, x_2, ..., x_n) in C-space. Motion planning finds a path for a robot in which the robot moves from a start position to a goal position without colliding with any obstacles in the environment. Motion planning has numerous applications including protein folding [1], computer-aided design (CAD) [4], robotic surgery [15, 21], and computer animation [13].

It is hard to find an exact motion planning solution due to the high dimensionality. Therefore, some randomized algorithms have been developed to address this issue, for example, sampling-based methods such as Probabilistic Roadmaps (PRMs) [12] and Rapidly-Exploring Random Trees (RRTs) [16]. Specifically, PRMs [12] generate random robot configurations to form a graph to represent C-space. PRMs, however, have been shown to perform poorly in generating samples in some difficult areas, e.g., narrow passages and obstacle surfaces [10] because the probability of generating a configuration is dependent on the volume of free space. The tendency is for oversampling to occur in relatively free parts of C-space and undersampling to occur in more obstructed parts of C-space, like the narrow passages. Therefore, some obstacle-based sampling methods have been proposed.

Obstacle-Based Probabilistic Roadmap (OBPRM) [2] is a specialized obstacle-based method which biases robot configurations to be close to the obstacle surfaces. OBPRM was used previously to predict ligand binding sites [3]. The work approximates ligands as linkage robots and proteins as obstacles and the result shows that OBPRM is able to accurately predict the ligand binding site since OBPRM is able to generate ligand configurations close to the binding site on the protein surfaces. With the aid of human input OBPRM performs even better.

Althouth OBPRM can generate samples close to the obstacle surfaces, the node distribution is biased by the shape of the obstacle and is unknown. Uniform configuration distribution around the obstacle surfaces is important because it requires the least amount of nodes to represent the C-space which can solve the motion planning problem faster.

UOBPRM (Uniform Obstacle-Based Probabilistic Roadmap) [25] , is a recently developed novel sampling method which guarantees a uniform node distribution around the obstacle surfaces and maintains efficiency at the same time.

1) UOBPRM: UOBPRM generates robot configurations in a uniform distribution around the obstacle surfaces. Algorithm 1 describes the node generation for UOBPRM. UOBPRM generates randomly distributed line segments with fixed length l by generating a random configuration c and extending it in a random direction $\overrightarrow{d}$. For each line segment, UOBPRM finds the intersection between the line segment and obstacle at a step size t. The valid configurations are kept as roadmap nodes.

UOBPRM finds the intersections between line segments and obstacles by checking the validity changes along the line segment. UOPBPRM generates intermediate configurations at a step size t for each line segment, and checks its validity. If the validities are different between two consecutive configurations, there is an intersection. The valid node is retained. The pseudocode for finding the intersection is shown in Algorithm 2. The validity checking is applied to the whole line segment since there can be multiple intersections on one line segment as the example shown in Figure 3.

Algorithm 1 UOBPRM Node Generation(n, l, t)

```
Input. A maximum attempts $n$, a line segment of length $l$, and
    a step size $t$
Output. A set of nodes $V$ uniformly distributed near obstacle
    surfaces
Set a bounding box whose margin is $l$ away from the
    obstacles (details on resetting the bounding box may be
    found in [25])
$V = \emptyset$
for $i = 1 \to n$ do
    Uniformly sample configuration $c$
    Generate a random direction $\overrightarrow{d}$
    Extend a segment $s$ from $c$ distance $l$ in direction $\overrightarrow{d}$
    $V \leftarrow$ Intersect$(s, l, t)$
end for
return $V$
```

III. Methods

A. Ligand and Protein Models

We model the ligand as a flexible, articulated-linkage robot and the protein as a rigid body. We also remove the hydrogen atoms in the ligands in order to simplify computation.

Algorithm 2 Intersect(s, l, t)

```
Input. A line segment $s$ of length $l$ and a step size $t$
Output. A set of intersections $I$
for $i = 1 \to (l/t)$ do
    Generate node $c_i$ along $s$
    if validity$(c_i) \neq$ validity$(c_{i+1})$ then
        Add the valid one to $I$
    end if
end for
return $I$
```

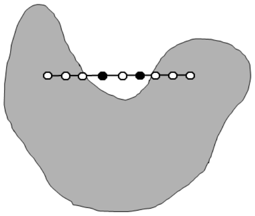


Fig. 3. Finding intersections between the line segment and the obstacle by checking the validity of intermediate configurations along segment. The valid node is retained at every validity change. Here, the valid nodes that are retained are solid.

B. Algorithm

To approximate binding affinity, we make the assumption that if the binding affinity is higher for a ligand sample, it is more likely to be buried deeper in the protein.

We use UOBPRM to generate ligand samples. For each ligand sample generated, we calculate the distance between the center of mass of the protein and the center of mass of the ligand sample. We find the minimum value amongst these distances. For the native structures of the protein-ligand complexes that are published in the literature, we also calculate the distance between the center of mass of the protein and the center of mass of the ligand. Next, we find the difference between the distance obtained from the native structure and the minimum distance for the samples generated by UOBPRM. The absolute value of this difference is then used as our approximated ligand binding affinity. Algorithm 3 describes the binding affinity calculation in this work.

Algorithm 3 Calculating Ligand Binding Affinity (P, L, n)

```
Input. A protein $P$, a ligand $L$, and the number $n$, of samples.
Output. An affinity score $a$, to represent the binding affinity
    between protein $P$ and ligand $L$.
for $i = 1 \to n$ do
    Generate UOBPRM samples $l_i$ for $L$
    $d_i \leftarrow$ distance(center of mass$(P)$, center of mass$(l_i)$)
    $min_L \leftarrow min\{d_i | \forall\ 1 \leq i \leq n\}$
end for
$d_{native} \leftarrow$ distance(center of mass$(P)$, center of mass$(L)$)
$a \leftarrow |min_L - d_{native}|$
return $a$
```

IV. Results

We implement our algorithm in C++ in a Linux 3.14 kernel with Fedora distribution version 20.

We obtain 2 sets of protein-ligand pairs from BindingDB [17] with known binding affinities that were previously determined experimentally. The first set contains one ligand and multiple proteins with different binding affinities and the second one has one protein with multiple ligands forming various complexes. For each protein-ligand pair in a set, we compute the binding affinity by the process described in Algorithm 3, which is used to rank the binding affinities among the protein-ligand pairs. Our simulation results are then compared to the published ranks.

We carry out experiments varying the line segment length l and number of samples n in UOBPRM. We start with an l value of 10 and an n value of 50 and then hold each of these values constant while varying the other value.

In more detail, for the set of one ligand and multiple proteins, we keep the line segment length l constant at 10 and determine our rank of binding affinities using various values of number of samples n of 50, 100, and 1000. Results are shown in Table I. We also determine our rank of binding affinities for this set keeping the number of samples constant at 50 and having different values for the line segment length of 5, 10, and 100 as shown in Table II. We repeat the above experimental setup for the set of one protein with multiple ligands and the results are shown in Tables III and IV. In Tables I and II our ranking matches the ranking from experimental binding affinities. For Table III when l is 10 and n is 50, our ranking does not match the experimentally-determined ranking. Likewise, for Table IV when l is 100 and n is 50, our ranking does not match the experimentally-determined ranking.

V. Discussion and Conclusion

In this paper, we propose a strategy which uses UOBPRM to study the ligand binding problem in computational biology. UOBPRM has been shown to generate ligand configurations close to protein surfaces. We experiment on 5 protein-ligand pairs with known binding affinities determined experimentally and find that UOBPRM is able to help rank binding affinity. This information may be useful to pharmaceutical companies in the area of computational or virtual drug screening.

Limitations of our approach include that we approximate the protein as a rigid body which may cause us to lose the true ligand binding site. When the true ligand binding site occurs deep in the protein, that space may be lost when the protein is transformed into a solid 3D model. At the same time, proteins can change conformation in reality when the ligand binds but we cannot capture this phenomenon when we model the protein as a rigid body. We have exceptions in our results that we are unable to explain within the current extent of our experiments. In the future, carrying out a wider range of varied parameters may be useful in order to be able to detect trends that may help to explain discordant results. In addition, we would like to increase the experimental set size above 5 protein-ligand pairs. We would also like to have protein-ligand pairs whose binding affinities differ by many more degrees of magnitude. This is because the discordant results may be due to the possibility that our approach is not sensitive enough to rank accurately protein-ligand pairs whose binding affinities differ by only one or two orders of magnitude. We plan to give the protein more flexibility. We also aim to consider other binding affinity measurements such as energy, rigidity, and compactness. Furthermore, we expect to vary the value of the step size t, in UOBPRM to evaluate its effect on the ranking we produce with our algorithm.

VI. Acknowledgments

Many thanks to Dr. Nancy Amato, my faculty mentor, for the opportunity to participate in this summer research program, for guidance, and for financial and other support; Dr. Shawna Thomas, post-doc, for her oversight; Hsin-Yi (Cindy) Yeh, my Ph.D. graduate student mentor, for her guidance, supervision, and support; and the Computing Research Association (CRA) and the Coalition to Diversify Computing (CDC) for funding this Distributed Research Experiences for Undergraduates (DREU) experience.

References

[1] N. M. Amato and G. Song. Using motion planning to study protein folding pathways. *J. Comput. Biol.*, 9(2): 149–168, 2002. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2001.

[2] N. M. Amato and Y. Wu. A randomized roadmap method for path and manipulation planning. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 113–120, 1996.

[3] O. B. Bayazit, G. Song, and N. M. Amato. Ligand binding with OBPRM and haptic user input: Enhancing automatic motion planning with virtual touch. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 954–959, 2001. This work was also presented as a poster at *RECOMB 2001*.

[4] H. Chang and T. Y. Li. Assembly maintainability study with motion planning. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 1012–1019, 1995.

[5] M. Dickson and J. P. Gagnon. The cost of new drug discovery and development. *Discovery Med.*, page June 20, 2009.

[6] P. Willett G. Jones and R. C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.*, 245:43–53, 1995.

[7] I. Halperin, B. Ma, H. Wolfson, and F. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, 2002.

[8] Matthew Herper. How much does pharmaceutical innovation cost? a look at 100 companies. *Forbes.com: Business section - Pharma and Healthcare*, August 2013.

TABLE I
ONE LIGAND AND MULTIPLE PROTEINS: VARIED NUMBER OF SAMPLES FOR CONSTANT L = 10

			Our Simulation ($n = 50$)		Our Simulation ($n = 100$)		Our Simulation ($n = 1000$)		Published
Ligand CID	Protein	Affinity (IC_{50})	Abs Δ Distance ($\mathring{A}$)	Rank	Abs Δ Distance ($\mathring{A}$)	Rank	Abs Δ Distance ($\mathring{A}$)	Rank	Rank
162204	2EGH	58 nM	0.2342	1	0.2932	1	1.8626	1	1
	4OOE	2390 nM	2.6459	2	4.5691	2	5.1965	2	2

TABLE II
ONE LIGAND AND MULTIPLE PROTEINS: VARIED LINE SEGMENT LENGTHS FOR CONSTANT N = 50

			Our Simulation ($l = 5$)		Our Simulation ($l = 10$)		Our Simulation ($l = 100$)		Published
Ligand CID	Protein	Affinity (IC_{50})	Abs Δ Distance ($\mathring{A}$)	Rank	Abs Δ Distance ($\mathring{A}$)	Rank	Abs Δ Distance ($\mathring{A}$)	Rank	Rank
162204	2EGH	58 nM	0.4082	1	0.2342	1	0.2942	1	1
	4OOE	2390 nM	1.6219	2	2.6459	2	4.5758	2	2

TABLE III
ONE PROTEIN AND MULTIPLE LIGANDS: VARIED NUMBER OF SAMPLES FOR CONSTANT L = 10

			Our Simulation ($n = 50$)		Our Simulation ($n = 100$)		Our Simulation ($n = 1000$)		Published
Protein	Ligand CID	Affinity (K_i)	Abs Δ Distance ($\mathring{A}$)	Rank	Abs Δ Distance ($\mathring{A}$)	Rank	Abs Δ Distance ($\mathring{A}$)	Rank	Rank
3W6H	768	5×10^{-4} nM	0.0695	1	0.0695	1	0.4146	1	1
	24530	12×10^{-4} nM	0.5548	2	1.4223	3	1.4908	2	2
	19366655	200×10^{-4} nM	1.2712	3	1.2712	2	1.6130	3	3

TABLE IV
ONE PROTEIN AND MULTIPLE LIGANDS: VARIED LINE SEGMENT LENGTHS FOR CONSTANT N = 50

			Our Simulation ($l = 5$)		Our Simulation ($l = 10$)		Our Simulation ($l = 100$)		Published
Protein	Ligand CID	Affinity (K_i)	Abs Δ Distance ($\mathring{A}$)	Rank	Abs Δ Distance ($\mathring{A}$)	Rank	Abs Δ Distance ($\mathring{A}$)	Rank	Rank
3W6H	768	5×10^{-4} nM	0.7803	1	0.0695	1	3.6673	3	1
	24530	12×10^{-4} nM	1.0039	2	0.5548	2	0.6711	1	2
	19366655	200×10^{-4} nM	1.5472	3	1.2712	3	1.8102	2	3

[9] Matthew Herper. The cost of creating a new drug now $5 billion, pushing big pharma to change. *Forbes.com: Business section - Pharma and Healthcare*, August 2013.

[10] D. Hsu, J-C. Latombe, and H. Kurniawati. On the probabilistic foundations of probabilistic roadmap planning. *Int. J. Robot. Res.*, 25:627–643, July 2006. ISSN 0278-3649.

[11] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott. Principles of Early Drug Discovery. *British journal of pharmacology*, 162(6):1239–1249, 2011.

[12] 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.

[13] Y. Koga, K. Kondo, J. Kuffner, and J.C. Latombe. Planning motions with intentions. In *Proc. ACM SIGGRAPH*, pages 395–408, 1995.

[14] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *JMB*, 161(2):269–288, 1982.

[15] Y. S. Kwoh, J. Hou, E. A. Jonckheere, and S. Hayati. A robot with improved absolute positioning accuracy for ct guided stereotactic brain surgery. *IEEE Trans. Biomed. Eng.*, 35(2):153–160, 1988.

[16] Steven M. Lavalle. Rapidly-exploring random trees: A new tool for path planning. Technical report, 1998.

[17] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*, 35:198–201, January 2007.

[18] T. Lengauer M. Rarey, B. Kramer and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.*, 261:470–489, 1996.

[19] Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.*, 16:2785–2791, 2009.

[20] Amit P. Singh, Jean-Claude Latombe, and Douglas L. Brutlag. A motion planning approach to flexible ligand binding. In *Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, pages 252–261, 1999.

[21] R. H. Taylor, B. D. Mittelstadt, H. A. Paul, W. Hanson, P. Kazanzides, J. F. Zuhars, B. Williamson, B. L. Musits, E. Glassman, and W. L. Bargar. An image-directed robotic system for precise orthopaedic surgery. *IEEE*

Trans. Robot. Automat., 10(3):261–275, 1994.

[22] M. Teodoro, G.N. Phillips, Jr, and L.E. Kavraki. Molecular docking: A problem with thousands of degrees of freedom. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 960–965, 2001.

[23] Shawna Thomas, Xinyu Tang, Lydia Tapia, and Nancy M. Amato. Simulating protein motions with rigidity analysis. *J. Comput. Biol.*, 14(6):839–855, 2007. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2006.

[24] Montserrat Vaque, Anna Ardrevol, Cinta Blade, M. Josepa Salvado, Mayte Blay, Juan Fernandez-Larrea, Lluis Arola, and Gerard Pujadas. Protein-ligand docking: A review of recent advances and future perspectives. *CURRENT PHARMACEUTICAL ANALYSIS*, 4(1):1–19, 2008.

[25] Hsin-Yi (Cindy) Yeh, Shawna Thomas, David Eppstein, and Nancy M. Amato. UOBPRM: A uniformly distributed obstacle-based PRM. In *Proc. IEEE Int. Conf. Intel. Rob. Syst. (IROS)*, pages 2655–2662, Vilamoura, Algarve, Portugal, 2012.