Stochastic Roadmap Simulation for The Study of Ligand-Protein Interactions

Mehmet Serkan Apaydin* **Carlos Guestrin*** **Chris Varma*** **Douglas L. Brutlag†**

Department of *Computer Science and †Biochemistry, Stanford University, Stanford, CA 94305
{apaydin,guestrin,varma}@cs.stanford.edu
brutlag@stanford.edu

Abstract

The dynamics of ligand-protein interactions are necessary to understand for the design of novel therapeutic agents. In this paper, we establish the use of Stochastic Roadmap Simulation (SRS) for the study of ligand-protein interactions through two studies. In our first study, we measure the effects of mutations on the catalytic site of a protein, a process called *computational mutagenesis*. In our second study, we focus on distinguishing the catalytic site from other putative binding sites. SRS compactly represents many Monte Carlo (MC) simulation paths in a compact graph structure, or roadmap. Furthermore, SRS allows us to analyze all the paths in this roadmap simultaneously. In our application of SRS to the domain of ligand-protein interactions, we consider a new parameter called escape time, the expected number of MC simulation steps required for the ligand to escape from the "funnel of attraction" of the binding site, as a metric for analyzing such interactions. Although computing escape times would probably be infeasible with MC simulation, these computations can be performed very efficiently with SRS. Our results for six mutant complexes for the first study and seven ligand-protein complexes for the second study, are very promising. In particular, the results agree well with the biological interpretation of the mutations, and the results show that escape time is a good metric to distinguish the catalytic site for five out of seven complexes.

1 Introduction

Understanding ligand-protein interactions is of fundamental importance. These interactions play a central role in biological processes essential to life. Studying these interactions is also an indispensable step in discovering new therapeutic molecules during the drug design process. Traditionally ligand-protein interactions are studied only through laboratory experiments, which are often time consuming and costly. With the rapid advances in molecular simulation techniques and computer hardware, computational methods become increasingly important for use in these studies. They compliment the laboratory methods by providing fast and inexpensive initial analyses to guide further examination through laboratory experiments.

In this paper, we study the binding affinity between ligands and proteins with a recently developed approach for analyzing molecular motion, called Stochastic Roadmap Simulation (SRS), a method for representing many MC simulations simultaneously in a compact graph structure [ABG$^+$02]. We introduce the notion of escape time as a measure of binding affinity. Intuitively escape time is the expected amount of time for a ligand to escape from the "funnel of attraction" in the binding site of a protein. Let us consider the conformation space of a ligand-protein complex with a suitably defined energy function. The binding site lies in a small region of the space. If the ligand is bound with high affinity, it would take much longer to escape from this region. We hypothesize that a longer escape time is a result of high energy barriers around the catalytic site, which are likely attributable to the energy distribution at the catalytic site. We thus examine the ligand-protein binding process with a dynamic model in which we directly account for the dynamics of the ligand in and around the catalytic site. In contrast, most previous computational approaches analyzing ligand-protein interactions, such as [MGH+98], employ static models and consider only the final bound conformation of the ligand, and they cannot be used to compute properties of the binding process.

In principle, escape time can be computed with simulation techniques. However, escape time is not a property of one molecular motion pathway, but an average property of many pathways. To estimate the escape time, we must calculate the time for a ligand to escape along many different pathways and then take the average. Classic simulation techniques such as the Monte Carlo (MC) and molecular dynamics methods are computationally intensive and impractical for this purpose on a workstation. The reason is that they focus on a single pathway at a time and are easily trapped in the local minima of the energy landscape. Therefore by combining SRS, which considers many pathways simultaneously and compactly, with tools from Markov Chain Theory, we can compute the escape time efficiently and achieve several orders of magnitude reduction in running time, compared with MC simulation.

To validate our approach, we conducted two studies and obtained very promising results. In the first study, we examined the effects of mutations at the catalytic site of a protein by performing computational mutagenesis. Specifically we mutated the residues near the catalytic site of lactate dehydrogenase and observed the effects of these mutations on the binding affinity by computing the escape time. In all six cases considered, our results are consistent with the biological interpretation of the mutations. Escape time can also serve as a useful discriminator for distinguishing the catalytic site from other putative binding sites, because the high energy barriers around the catalytic site lead to larger escape time. So, in the second study, we computed escape times at different binding sites of seven ligand-protein complexes. In five of the seven cases, the escape time clearly distinguishes the catalytic site with value at the catalytic site over two orders of magnitude larger than those at other potential binding sites. In all the studies, the escape times were computed within a few minutes per binding site on a desktop computer. These results show that our approach provides an efficient computational tool for investigating ligand-protein interactions.

The rest of the paper is organized as follows. In Section 2, SRS is defined and its application to ligand-protein interactions is outlined. The efficient computation of escape times with SRS is presented in Section 3. Section 4 outlines the modeling decisions for the ligand-protein studies. The computational mutagenesis study is presented in Section 5. The second study, distinguishing catalytic sites from other potential binding sites is described in Section 6.

2 Stochastic Roadmap Simulation for ligand-protein interactions

Simulation techniques, such as Monte Carlo Simulation, or Molecular Dynamics, can be used to study ligand-protein interactions. For example, one can suggest modes of binding of a ligand to a catalytic site, and understand the role of catalytic residues in binding, using molecular dynamics [NSMcC01]. However, the stochastic nature of the process requires one to gather many simulation paths to make such study thorough and precise. Monte Carlo Simulation and Molecular Dynamics are limited to generating one simulation trajectory at a time, thus making it impractical.

Stochastic Roadmap Simulation(SRS) [ABG+02] has been proposed as an efficient and accurate simulation tool to study molecular motion. SRS constructs a roadmap, which is a discrete representation of molecular motion. A roadmap contains many Monte Carlo simulation paths simultaneously and enables one to perform key computations efficiently. Previously, SRS was applied to the computation of transmission coordinate(pfold) in protein folding, demonstrating orders of magnitude speedup and better accuracy in the computation of the pfold parameter.

In using SRS, one needs to first represent the ligand and protein complexes studied. The conformation of a ligand and the associated protein can be specified in various ways, such as by the 3D coordinates of one of the atoms and the torsional angles for the conformation of the ligand and the backbone torsional angles (ϕ and ψ) for the protein. Formally, a conformation of d parameters is specified by a vector $(\theta_1, \theta_2, \ldots, \theta_d)$. The set of all possible conformations form the *conformation space* $\mathcal{C}$. A point in $\mathcal{C}$ corresponds to a particular assignment to the parameters that specify the conformation of both the ligand and the protein. The conformational parameters determine the interaction between atoms of the molecules and between the molecules and the medium, *e.g.*, the van der Waals and electrostatic forces. These interactions give rise to the attractive and repulsive forces that dictate the motion of molecules. SRS assumes that the interactions are described by an energy function $E(q)$, which depends only on the conformation q of the molecules; it does not require E to have any particular properties or functional forms.

A pathway in $\mathcal{C}$ corresponds to a particular motion of the ligand and the associated protein. SRS encodes many such pathways in $\mathcal{C}$ with a directed graph G, called a roadmap. Each node of the roadmap G is a randomly sampled conformation in $\mathcal{C}$. Each (directed) edge between two nodes v_i and v_j carries a weight P_{ij}, which is the probability for the molecules to transition from v_i to v_j. The probability P_{ij} is 0 if there is no edge between v_i and v_j. Otherwise, the value of P_{ij} depends on the energy difference between v_i and v_j. SRS thus adopts a stochastic view of molecular motion: P_{ij} represents the probability that the molecules will next move to conformation v_j, given that they are currently in v_i. To construct the roadmap, the algorithm samples n conformations independently at random from $\mathcal{C}$. More specifically, for each node v_i, each conformational parameter $\theta_i, i = 1, 2, \ldots$ is sampled from its allowable range according to some chosen distribution. For every node v_i, one then finds the k nearest neighbors of v_i, according to a suitable metric such as the RMS or Euclidean distance in $\mathcal{C}$. Let N_i denote the set of k nearest neighbors of v_i. The algorithm then computes the transition probability P_{ij} between every pair of neighboring nodes v_i and v_j, where v_j is in N_i. P_{ij} is computed based on $\Delta E_{ij} = E(v_j) - E(v_i)$, the energy difference between the conformations v_i and v_j. In formula,

$$P_{ij} = \begin{cases} (1/|N_i|)\exp(-\Delta E_{ij}/k_{\mathrm{B}}T), & \text{if} \Delta E_{ij} > 0; \\ 1/|N_i|, & \text{otherwise}; \end{cases}$$

where k_{B} is the Boltzmann constant, T is the temperature, and $|N_i|$ is the number of neighbors of node v_i, excluding itself. If a node v_j is not in N_i, then v_i and v_j are too far apart for their energy difference to be a good basis for estimating the transition probability, and we set $P_{ij} = 0$. Finally the self-transition probabilities are defined as:

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij},$$

which ensures that the transition probabilities from any node sum up to 1.

Typically, Monte Carlo simulation generates random paths through $\mathcal{C}$ corresponding to potential motions of the ligand and the protein. Such paths are interesting for understanding the energy landscape and exploring the kinetics of molecular motion, as well as determining binding sites (see, *e.g.*, [Fer99, KS96]).

However, Monte Carlo simulation focuses on only one pathway at a time and easily gets stuck in the local minima of the energy function, repeatedly sampling many similar conformations without obtaining much new information. SRS constructs a roadmap containing many Monte Carlo simulation paths by sampling directly from the space of all pathways. It processes these paths together using algebraic methods, thus greatly reducing computation time [ABG$^+$02]. Furthermore, the computation does not suffer from the local-minima problem encountered in Monte Carlo simulation.

3 Escaping from a putative binding site

A protein contains many cavities of low energy where a ligand could potentially bind. We refer to these locations as *putative binding sites*. An interesting measure of affinity of a ligand to a putative binding site could be the expected "amount of time" a ligand would take to escape the "funnel of attraction" of this site. We call such quantity the *escape time*. To obtain a precise definition of the escape time, we must formalize two notions: "funnel of attraction" and "amount of time". The notion of funnel of attraction of a putative binding site has been used in other studies of ligand-protein interaction, such as [CV01], where the funnel is defined as all ligand conformations within 10 Å RMSD of the bound conformation. In this section, we will use a generic definition, where each putative binding conformation v is associated with a set of bound conformations $\mathcal{F}$. If we use this definition, then $\mathcal{F}$ would correspond to all conformations within 10 Å RMSD of v. Next, we must make precise the idea of "amount of time". If we are using a simulation method, such as MC simulation, then we can define amount of time as the number of simulation steps. Thus, we can now make precise the definition of escape time:

Definition 1 *The* escape time τ *from a putative binding site* v *is the* expected *number of MC simulation steps, starting from* v*, required for the ligand to reach a conformation outside the funnel of attraction* $\mathcal{F}$ *of* v. ∎

A naive approach for computing escape times is to run many MC simulations, starting from v and to average the number of steps each simulation took to reach a conformation outside $\mathcal{F}$. However, this approach is quite impractical due to the computation cost of running many MC simulations. On the other hand, a roadmap G compactly encodes many MC simulation paths. An insight resulting from our choice of transition probabilities is that a roadmap implicitly defines a Markov chain that captures the stochastic nature of molecular motion. This allows us to take advantage of powerful tools from the Markov chain theory [ABG+02]. In the remainder of this section, we focus on how one such tool, first-step analysis, can be applied to compute escape times by considering all paths on the roadmap simultaneously. Consider a roadmap G representing the motion of a ligand-protein complex. Let v_i be a node on the roadmap corresponding to a potential bound conformation and $\mathcal{F}^{(i)}$ be a set of nodes in G that lie in the funnel of attraction of v_i, *e.g.*, all nodes on the roadmap within 10 Å RMSD of v_i. Now, suppose that we are interested in finding the escape time τ_i starting from v_i. The naive approach to compute τ_i would be to perform many simulation runs on the roadmap, starting from v_i and average the number of steps taken by each run to obtain an estimate of τ_i. This approach has a high variance in the estimate, due to the stochastic nature of the simulation procedure and thus requires a large number of simulation runs in order to achieve reasonable results. In contrast, first-step analysis computes τ_i without the need for explicit simulation. First-step analysis proceeds by conditioning on what happens after the first step of simulation. Suppose that we start at some node $v_i \in \mathcal{F}^{(i)}$ and perform one transition step. First τ_i is increased by one. Then, in the next step, we reach either a state outside the funnel $\mathcal{F}^{(i)}$ or another node $v_j \in \mathcal{F}^{(i)}$. In the former case, we simply stop as the ligand has escaped. In the latter case, the expected number of steps from then on is exactly the escape time starting from v_j for funnel $\mathcal{F}^{(i)}$, given by τ_j. More formally, we have the following system of self-consistent equations:

$$\tau_i = 1 + \sum_{v_j \notin \mathcal{F}^{(i)}} P_{ij} \cdot 0 + \sum_{v_j \in \mathcal{F}^{(i)}} P_{ij} \cdot \tau_j, \quad \text{for every } v_i \in \mathcal{F}^{(i)}. \tag{1}$$

In the second term of (1), P_{ij} is multiplied by zero, because we the simulation is completed as soon as the ligand escapes. See Figure 1 for an illustration. The linear system in (1) contains one equation and one unknown for each node v_i in $\mathcal{F}^{(i)}$. A unique solution to (1) is guaranteed to exist, because the roadmap G contains only one strongly-connected component by construction, and so the Markov chain represented by G is ergodic [TK94]. By solving the linear system algebraically, we obtain τ_j for all the nodes simulta-

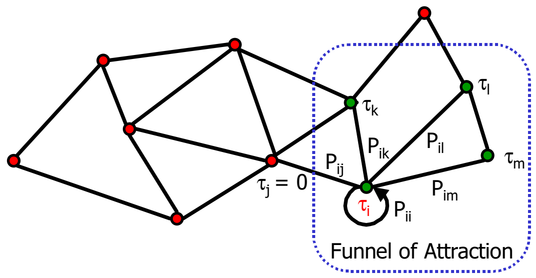

Figure 1: Computing escape time τ_i from node v_i.

neously, without any explicit simulation. In particular, we obtain the escape time τ_i for the bound conformation v_i.

4 Ligand-protein modeling

In our study, we represented the ligand-protein complexes as in [SLB99, ASBL01]. The protein was assumed as rigid, whereas the ligand was flexible. One atom in the ligand was assumed to be the base and was assigned 5 DOFs, whereas each additional non-terminal atom was associated with a torsional DOF. The bonds in a ring were assumed as rigid and were assigned no DOFs. The bond angles and lengths were assumed constant. To calculate the energy of interaction between the ligand and the protein, we used a potential function that incorporates electrostatic and van der Waals components as well as solvation free energies as approximated by continuum models [SH94]. A dielectric of 80 was used to model the solvent and a dielectric of 1 was used to model the solute as is consistent with previous methods [RK00]. We use the Delphi program [SH90] which employs the Poisson-Boltzmann equation to calculate the electrostatic and solvation free energy terms of our potential function. We thus calculated an electrostatic potential grid at a resolution of either 1A or 0.5A. As in previous work [SLB99], Van der Waals potentials were computed on the same grid. The energy of the ligand was also calculated as in prior work [SLB99], and the charges on each atom of each ligand were computed as a formal charge taking into account resonance structures of the molecule. In our study, we defined the funnel of attraction of a potential binding site as in [CV01], i.e., the set of conformations within 10A rmsd of the bound ligand complex. We also repeated our experiments with funnels of 6A and 8A radii, obtaining comparable results.

5 Analyzing the effects of mutations

We first applied SRS to the analysis of the effects of mutations in the catalytic site of a protein on the escape time of a ligand.

5.1 Computational mutagenesis

Computational mutagenesis is a new and exploratory area of computer-aided protein design. Computational mutagenesis is based on the biological method of site-directed mutagenesis. In this method, a few amino acids are either deleted entirely or replaced by other amino acids, or the side chains of the amino acids are altered. Site-directed mutagenesis has proven quite useful for many studies, including substrate recognition and identification of catalytic amino acids [CWC+86]. The mutations made through this method are specific in terms of what changes are made, local in terms of exactly which amino acids are affected, and sound in terms of having no significant structural ramifications. Computational mutagenesis embodies these concepts from site-directed mutagenesis, but enables mutations to be performed *in silico* providing the obvious benefits of speed and ease at perhaps the expense of model accuracy. Reyes and Kollman, for example, have shown encouraging early results in utilizing computational mutagenesis to study binding specificity [RK00].

5.2 Mutagenesis study on lactate dehydrogenase

We employed computational mutagenesis in order to study the sensitivity of SRS when applied to ligand-protein interactions by detecting escape times of a ligand from its target. In particular, we used oxamate (an inactive analogue of pyruvate) and lactate dehydrogenase.

Lactate dehydrogenase (LDH) LDH is an enzyme that when bound to its coenzyme NADH is

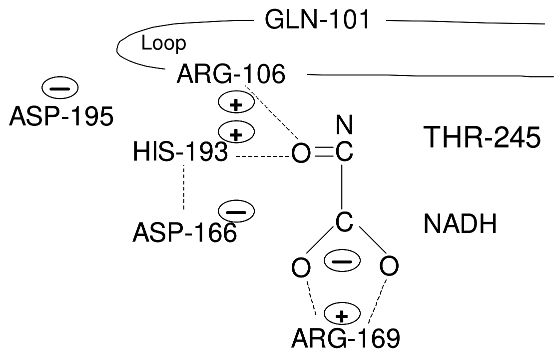

Figure 2: The chemical environment of LDH-NADH-substrate complex. Hydrogen atoms are not explicitly modeled.

able to catalyze the reduction of pyruvate to lactate. LDH is a well-studied enzyme [CWHH85, Har89]. In fact, LDH has been proposed as a general framework on which to design and synthesize new enzymes [DWH+91]. We use dogfish apo-lactate dehydrogenase (PDB: 1LDM) as a model on which to perform computational mutagenesis.

The active site of LDH is well understood. The chemical environment of oxamate in its bound conformation in the LDH-NADH-substrate complex is shown in Figure 2. The amino acids that play a significant role in the catalytic activity of the enzyme are shown. Arg169 assists in orienting and binding the substrate [HCW+87]. Arg106 polarizes the carbonyl bond on the substrate [CWC+86]. His193 is an important catalytic residue, which donates a proton to the substrate during its reduction [HLSR75]. His193 is then stabilized by Asp166 [CBA+88]. In native LDH, before the binding of the coenzyme or the substrate, a loop of polypeptide chain (residues 97 to 107) is positioned away from the active site. After the binding of coenzyme and the substrate, a rearrangement in protein structure is induced which results in the loop being positioned over the active site as shown in Figure 2.

Mutations Two sets of mutations were performed on LDH based largely on prior *in vitro* work [DWH+91]. The first set included changing of charged and catalytic amino acids (His193 $\rightarrow$ Ala, Arg106 $\rightarrow$ Ala, and both His193 $\rightarrow$ Ala and Arg106 $\rightarrow$ Ala). These mutants cause a large reduction in the energetic structure of the active site, thus, can provide insights into the sensitivity of SRS to coarse changes in the system. The second set of mutants (Asp195 $\rightarrow$ Asn, Gln101 $\rightarrow$ Arg, Thr245 $\rightarrow$ Gly) play a cursory role in catalysis and thus were expected to have a less significant effect. These mutants, on the other hand, can provide us with insights into the sensitivity of SRS to fine changes in the system, as they cause small or no reduction in the energetic structure of the active site.

Mutations were performed using Sybyl (distributed by Tripos Inc.). No structural recalculation or minimization was performed as it was assumed that the structural change upon mutation is insignificant, as in other computational mutagenesis work [RK00]. The energy potential grid and escape times were then calculated as described in Section 3 and Section 4. The roadmaps generated contained 4000 nodes sampled over the whole conformation space and 100 extra nodes sampled within 10 Å RMSD of the bound conformation. Furthermore, several other sampling schemes were employed which corroborated our described results–thus attesting to the robustness of the method for such work.

5.3 His193 $\rightarrow$ Ala

His193 is an important catalytic and charged amino acid. Replacing His193 with Ala would cause a significant reduction in the energetic structure of the active site [WHF+88], which results in less tight binding between enzyme and substrate. Therefore, decreasing the affinity of the substrate for the enzyme. We would expect a faster escape from the bound conformation. (see Table 1).

5.4 Arg106 $\rightarrow$ Ala

Arg106 is also an important catalytic and charged amino acid. Similar to His193, we would expect a significant reduction in the energetic structure of the active site [WHF+88], which would lead to a reduced affinity between enzyme and substrate. Thus, the substrate would be able to escape in less time from the bound conformation

Mutant	Bound Energy (kcat/mol)	Escape Time	Expected Affect
Wild type	0.233467	3.216e+06	N/A
His193 $\rightarrow$ Ala and Arg106 $\rightarrow$ Ala	4.526738	4.126e+02	His193 and Arg106 are the most important catalytic residues which provide much of the positive charge; we would expect a significant decrease in escape time.
His193 $\rightarrow$ Ala	-1.370748	3.381e+03	His193 is an important catalytic residue which provides some of the positive charge; we would expect a decrease in escape time.
Arg106 $\rightarrow$ Ala	1.305369	2.550e+02	Arg106 is an important catalytic residue which provides some of the positive charge; we would expect a decrease in escape time.
Asp195 $\rightarrow$ Asn	-9.258782	5.221e+07	Asp195 likely plays a significant role in charge conservation by providing a negative charge; we would expect a noticeable increase in escape time.
Gln101 $\rightarrow$ Arg	-8.516694	1.669e+06	Gln101 plays an important role in loop closure; we would expect not to see an affect on escape time as the protein is held rigid in our experiments.
Thr245 $\rightarrow$ Gly	-6.628186	4.607e+05	Thr245 employs a large side chain and thus reduces the total size of the active site-Gly is much smaller; we would expect a noticeable decrease in escape time.

Table 1: Effects of mutations on the catalytic site. Escape times are the geometric average of 20 roadmaps. Results agree well with biologically expected values.

when compared to wild type.

5.5 His193 $\rightarrow$ Ala and Arg106 $\rightarrow$ Ala

Both His193 and Arg106 are necessary catalytic and charged amino acids for enzymatic function of LDH. Thus, their replacement with Alanine would result in a significant reduction in energetic structure of the chemical environment of the LDH-substrate-complex [WHF$^+$88]. Therefore, we would expect the substrate to quickly escape from the active site.

5.6 Asp195 $\rightarrow$ Asn

Asp195 likely plays a significant role in charge conservation by providing a negative charge. Thus, its replacement with the neutral Asn would likely affect the energetic structure of the active site [WHF$^+$88] by increasing the affinity of the substrate for the active site. This would result in slower escape for the substrate.

5.7 Gln101 $\rightarrow$ Arg

Gln101 plays an important role in loop movement [WHF$^+$88]. Recall that binding of NADH and substrate induces a conformational change on the loop region causing it to close over the active site. Gln101 is replaced by Arg which is a positively charged amino acid, however, the location of the mutation is on the outside of the loop, therefore the additional charge can be assumed to be negligible when computing escape time. Furthermore, since our LDH is held rigid in these experiments, the Gln101 $\rightarrow$ Arg mutation is not expected to cause significant change in escape times.

ligand	protein	num random nodes	num DOFs
oxamate	1ldm	8000	7
streptavidin	1stp	8000	11
hydroxylamine	4ts1	8000	9
COT	1cjw	8000	21
THK	1aid	8000	9
IPM	1ao5	8000	10
PTI	3tpi	8000	13

Table 2: Roadmap parameters.

5.8 Thr245 $\rightarrow$ Gly

Thr245 employs a large side chain and thus reduces the total volume of the active site. In order to increase the volume of the active site without causing significant energetic restructuring of the active site, Thr245 was replaced by Gly, which has a much smaller side chain resulting in a net increase in total volume of the active site [WHF$^+$88]. Thus, escaping should become easier for the substrate.

6 Distinguishing the catalytic site

The catalytic site is the location on the protein surface where the ligand binds and performs its activity. There is shape and electrostatic complementarity between the catalytic site and the ligand, which enables a tightly bound ligand-protein complex. There has been work on using this complementarity to predict the location of the catalytic site, such as [NW99]. In this study, we focused on whether we can distinguish the catalytic site from a set of putative binding sites using escape time computations, rather than attempting to find the catalytic site by random sampling. Therefore, we added the catalytic site to the roadmap, and computed the escape time from regions around the catalytic site, as well as from the region around other putative binding sites. In [SLB99], 3 ligand-protein complexes were studied. For these complexes, the energy of the bound state was found to be an unsatisfactory discriminator between the catalytic site and other putative binding sites. Instead, another metric, the average path weight of the paths entering and leaving the putative binding sites was introduced. By considering the energetically most feasible paths entering and leaving the catalytic site, it was suggested that there is an energy barrier around the catalytic site. This high average path weight increased the difficulty for the ligand to enter and leave the catalytic site. On the other hand, with SRS we can consider all the possible molecular pathways, instead of only the energetically most feasible ones. In addition, with first step analysis, we can compute analytically the average number of simulation steps the ligand stays within the funnel of attraction of some potential binding site. Thus, measuring the effect of the whole energy barrier, rather than just a small part corresponding to the most feasible path. Furthermore, our results can be more precisely interpreted, as the escape time corresponds to the number of MC simulation steps required to escape the funnel of attraction of the putative binding site. For our study, we considered not only the 3 complexes of [SLB99], but a total of 7 ligand-protein complexes. These complexes and roadmap information are listed in Table 2.

protein	bound state	site 1	site 2	site 3	site 4
1ldm	-11.79	-13.57	-11.78	-11.38	-11.24
1stp	-15.06	-14.35	-13.52	-12.42	-12.21
4ts1	-19.44	-14.61	-14.60	-13.60	-12.30
1cjw	-11.69	-18.02	-15.24	-14.13	-14.11
1aid	-11.23	-22.17	-17.65	-15.81	-15.46
1ao5	-7.45	-13.13	-11.04	-10.83	-9.52
3tpi	-25.19	-15.99	-13.66	-13.16	-12.61

Table 3: Energy values corresponding to the bound state and putative binding sites, in kcal/mol.

First, putative binding sites were selected by presampling 10,000 random nodes in the landscape, and then performing random descent starting from conformations of lowest energy. In addition to the true binding site, the four conformations that have the lowest energies, that are close to the protein surface (less than 5Å RMSD) and distant from each other (greater than 10Å) were selected as the putative binding sites.

For each complex, 20 roadmaps were generated. Each roadmap was composed of a set of random conformations sampled as described in Section 4 (the number of conformations is given in Table 2, third column). In addition, 100 extra conformations were sampled within 10 Å of each putative binding site, as in [SLB99].

Table 3 shows the energy values corresponding to the catalytic and the putative binding sites. Note that the catalytic site is not the lowest energy conformation for 1ldm, 1cjw, 1aid and 1ao5. Thus, we observe, as in [SLB99], that the energy of a conformation is not a good criteria for distinguishing the catalytic site. Table 4 shows the resulting escape times from each putative binding site, obtained using 20 roadmaps for each ligand-protein complex. The roadmaps containing more than one connected component were discarded. The results presented are the geometric average of the remaining roadmaps. The system was implemented in two parts: A C++ part for generating the roadmaps and for energy computations, which required less than 4 minutes of computing time for a roadmap of 8000 nodes utilizing a Pentium III 800 MHz 1GB RAM workstation;

protein	bound state	site 1	site 2	site 3	site 4
1ldm	7.4e+05	2.0e+06	5.3e+05	5.6e+05	1.9e+05
1stp	3.4e+09	1.1e+07	4.2e+05	2.5e+05	9.3e+04
4ts1	2.9e+11	2.0e+06	1.2e+06	7.6e+05	8.9e+04
1cjw	1.1e+09	5.3e+06	2.6e+04	1.0e+04	3.2e+04
1aid	5.4e+06	2.4e+08	3.1e+06	1.3e+05	9.9e+04
1ao5	1.3e+09	5.3e+06	2.3e+05	1.4e+05	8.3e+03
3tpi	2.9e+11	9.2e+05	8.2e+04	3.2e+04	1.2e+04

Table 4: Escape time from regions around putative binding sites

and a Matlab portion, which computed the escape times, requiring less than 4.5 minutes to find the escape time for the 5 sites for the same roadmap.

These results illustrate that in 5 out of 7 complexes, the escape time from the funnel around the catalytic site is larger than any other escape time by at least an order of magnitude. For most cases, the differences were of at least two orders of magnitude, illustrating the intuition that escape time is a good metric for distinguishing the catalytic site.

7 Discussion

In this paper, we applied Stochastic Roadmap Simulation (SRS) to the analysis of ligand-protein interactions. In our studies, we considered the escape time, the expected number of MC simulation steps required for the ligand to escape from the funnel of attraction of the binding site, as a metric for analyzing ligand-protein interactions. Although computing escape times would probably be infeasible with MC simulation, with SRS, these computations can be performed very efficiently. In the first study, we provide evidence of SRS high sensitivity through its ability to detect changes in the chemical environment in ligand-protein interaction studies. Specifically, we measured the effects of mutations on the catalytic site of a protein, a process called computational mutagenesis. For six mutations of a protein, we computed the escape times of the ligand in the bound conformation and compared these quantities to the wild type. Our simulation results are very promising. In all cases, the escape time results obtained with SRS agreed with the biological interpretation of the mutation, establishing the sensitivity of SRS for use in ligand-protein interactions. In our second study,we established the employment of SRS for the study of ligand-protein interactions by calculating escape time as a metric for distinguishing the catalytic site from four other putative binding sites. In five out of seven complexes, escape time was a good metric, distinguishing the catalytic site by over two orders of magnitude from the other putative binding sites. Although, in our preliminary study escape time does not always distinguish the catalytic site from other binding sites, it seems to be a very strong indicator that could be used in conjunction with domain knowledge or other tools. Furthermore, this study makes clear the applicability of SRS to ligand-protein interaction studies.

Acknowledgments: This work has been partially funded by an NSF-ITR grant and a grant from Stanford's Bio-X program. Apaydin was supported by the D.L. Cheriton Stanford Graduate Fellowship. Brutlag was supported National Human Genome Research Institute grant HGF02235. Guestrin was supported by a Siebel Scholarship and by the Sloan Foundation. This paper has greatly benefited from discussions with D. Koller, J.C. Latombe and A. Singh. In addition, we thank D. Hsu for many invaluable comments.

References

[ABG+02] M. S. Apaydin, D. Brutlag, C. Guestrin, D. Hsu, and J. C. Latombe. Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. In *In the Sixth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, Washington D.C., April 2002.

[ASBL01] M. S. Apaydin, A. Singh, D. Brutlag, and J. C. Latombe. Capturing molecular energy landscapes with probabilistic conformational roadmaps. In *IEEE International Conference on Robotics and Automation (ICRA-01)*, Seoul, May 2001.

[CBA+88] A. Clarke, H. Wilks D. Barstow, T. Atkinson, W. Chia, and J. Holbrook. An investigation of the contribution made by the carboxylate group of an active site histidine-aspartate couple to binding and catalysis in lactate dehydrogenase. *Biochemistry*, 27:1617 – 1622, 1988.

[CHR95] M. Cummings, T. Hart, and J. Read. Monte carlo docking with ubiquitin. *Prot. Science*, 4(5):885 – 899, 1995.

[CV01] C.J. Camacho and S. Vajda. Protein docking along smooth association pathways. *PNAS*, pages 10636 – 10641, 2001.

[CWC+86] A. Clarke, D. Wigley, W. Chia, D. Barstow, T. Atkinson, and J. Holbrook. Site-directed mutagenesis reveals the role of a mobile arginine residue in lactate dehydrogenase catalysis. *Nature*, 324:699 – 702, 1986.

[CWHH85] A. Clarke, A. Waldman, K. Hart, and J. Holbrook. The rates of defined changes in protein structure during the catalytic cycle of lactate dehydrogenase. *Biochim. Biophys. Acta*, 829:397 – 407, 1985.

[DWH+91] C. Dunn, H. Wilks, D. Halsall, T. Atkinson, A. Clarke, H. Muirhead, and J. Holbrook. Design and synthesis of new enzymes based on the lactate dehydrogenase framework. *Phil. Trans. R. Soc. Lond.*, 332:177 – 184, 1991.

[Fer99] A. Fersht. *Structure and Mechanism in Protein Science: A guide to Enzyme Catalysis and Protein Folding*. W. H. Freeman and Company, New York, 1999.

[Har89] K. Hart. *An investigation of the molecular basis of substrate specificity in lactate dehydrogenase*. Ph.d. thesis, University of Bristol, 1989.

[HCW+87] K. Hart, A. Clarke, D. Wigley, A. Waldman, W. Chiaand D. Barstow, T. Atkinson, J. Jones, and J. Holbrook. A strong carboxylate-arginine interaction is important in substrate orientation and recognition in lactate dehydrogenase. *Biochim. Biophys. Acta*, 914:294 – 298, 1987.

[HLSR75] J. Holbrook, A. Liljas, S. Steindel, and M. Rossmann. Lactate dehydrogenase. *Enzymes*, 11a:191 – 293, 1975.

[KS96] A. Kolinski and J. Skolnick. *Lattice Models of Protein Folding, Dynamics and Thermodynamics*. Chapmann and Hall, New York, 1996.

[Lea96] A. Leach. *Molecular modelling: principles and applications*. Prentice Hall, New York, 1996.

[MGH+98] G. Morris, D. Goodsell, D. Halliday, R. Huey, W. Hart, R. Belew, and A. Olson. Automated docking using lamarckian genetic algorithm and an empirical binding free energy function. *J. Comp. Chem.*, 19:1639 – 1662, 1998.

[NSMcC01] H. Ni, C. A. Sotriffer, and J. A. McCammon Ordered water and ligand mobility in the HIV-1 Integrase-5 CITEP complex: A molecular dynamics Study *J. Med. Chem.*, 44:3043–3047, 2001.

[NW99] R. Nussinov and H. Wolfson. Efficient computational algorithms for docking, and for generating and matching a library of functional epitopes i. rigid and flexible hinge-bending docking algorithms. *Combinatorial Chemistry and High Throughput Screening*, 2:277–287, 1999.

[RK00] C. Reyes and P. Kollman. Investigating the binding specificity of u1a-rna by computational mutagenesis. *J. Mol. Biol.*, 295:1 – 6, 2000.

[SA01] G. Song and N. Amato. Using motion planning to study protein folding pathways. In *Proceedings the 5th International Conference on Computational Molecular Biology (RECOMB)*, pages 287–296, Montreal, April 2001.

[SH90] K. Sharp and B. Honig. Electrostatic interactions in macromolecules: theory and applications. *Ann Rev Biophys Chem.*, 19:301 – 332, 1990.

[SH94] K. Smith and B. Honig. Evaluation of the conformational free energies of loops in proteins. *Protein: Struct. Funct. Genet.*, 18:119 – 132, 1994.

[SLB99] A. P. Singh, J. C. Latombe, and D. L. Brutlag. A motion planning approach to flexible ligand binding. In *International Conference on Intelligent Systems for Molecular Biology*, pages 252–261, Heidelberg, 1999.

[TK94] Howard Taylor and Samuel Karlin. *An Introduction to Stochastic Modeling*. Academic Press, New York, 3rd edition, 1994.

[WHF+88] H. Wilks, K. Hart, R. Feeney, C. Dunn, H. Muirhead, W. Chia, D. Barstow, T. Atkinson, A. Clarke, and J. Holbrook. A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *science*, 242:1541 – 1544, 1988.