

CS 273
Prof. Serafim Batzoglou
Prof. Jean-Claude Latombe
Spring 2006

Molecular Motion Pathways:
Computation of Ensemble Properties with Probabilistic Roadmaps
(Lecture 17; tables and figures are taken from lecture slides)

Lecturer: Prof. Jean-Claude Latombe

Scribe: Chioma Osondu

1. Introduction

In the past seven years, there has been work done on using the robotics technique of probabilistic roadmaps to predict ensemble properties of molecules. The different approaches to predicting ensemble properties using probabilistic roadmaps discussed below are drawn from the following papers:

- A.P. Singh, J.C. Latombe, and D.L. Brutlag. A Motion Planning Approach to Flexible Ligand Binding. *Proc. 7th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, AAAI Press, Menlo Park, CA, pp. 252-261, 1999.
- N.M. Amato, K.A. Dill, and G. Song. Using Motion Planning to Map Protein Folding Landscapes and Analyze Folding Kinetics of Known Native Structures. *J. Comp. Biology*, 10(2):239-255, 2003.
- M.S. Apaydin, D.L. Brutlag, C. Guestrin, D. Hsu, J.C. Latombe, and C. Varma. Stochastic Roadmap Simulation: An Efficient Representation and Algorithm for Analyzing Molecular Motion. *J. Comp. Biology*, 10(3-4):257-281, 2003.
- N. Singhal, C.D. Snow, and V.S. Pande. Using Path Sampling to Build Better Markovian State Models: Predicting the Folding Rate and Mechanism of a Tryptophan Zipper Beta Hairpin. *J. Chemical Physics*, 121(1): 415-425, 2004.
- J. Cortés, T. Siméon, M. Renaud-Siméon, and V. Tran. Geometric Algorithms for the Conformational Analysis of Long Protein Loops. *J. Comp. Chemistry*, 25:956-967, 2004.

One example of ensemble properties is the probability of folding P_{fold} ¹. Given the denatured (unfolded) state of a protein and its native (folded) state; for any new configuration of the protein, P_{fold} is the measure of the kinetic distance to the given folded state. $1 - P_{fold}$ is the measure of the kinetic distance to the unfolded state. P_{fold} is large when the new conformation is relatively close to the folded state and small when the new conformation is closer to the unfolded state. P_{fold} is an ensemble property because it does not define a single pathway to either the folded state, so multiple motion pathways are allowed as long as they arrive at the folded state of the protein. Other examples of ensemble properties include, for folding: order of formation of SSE's, folding rate / mean first passage time and key intermediates; for binding, average time to escape from binding site and average energy barrier.

2. Motivation

Molecular motion is an essential process of life. This is evident in the fact that certain diseases, for example, mad cow disease, are caused by misfolding. Also, drug molecules can act as

inhibitors by binding to active sites of proteins. It is therefore necessary to gain a better understanding of molecular motion.

Unfortunately, very few experimental tools such as the Nuclear Magnetic Resonance (NMR) spectrometer are available for this task. Using the NMR spectrometer, it is possible to observe some dynamic properties of proteins, unlike x-ray crystallography.

Computer simulations using Monte Carlo (MC) simulations and molecular dynamics (MD) could also be used. This method however has two main drawbacks:

1. Each simulation run yields a single pathway, while molecules tend to move along many different pathways. This does not meet the requirement that properties observed are indeed ensemble properties.
2. Each simulation run wastes too much time exploring local minima, which makes this an expensive process.

3. Roadmap-Based Representation

This is a compact way of representing many motion pathways because it encodes an enormous number of pathways. However, it is coarse compared to MC and MD simulations. Also, efficient algorithms for analyzing multiple pathways can be borrowed from roadmaps for robot motion planning.

3.1. Application of Roadmaps to Ligand-Protein Binding (Initial Work)

This is a brief description of the work done by A.P. Singh, J.C. Latombe and D.L. Brutlag, which was a study of ligand-protein binding. Actual details can be found in the paper referenced in the introduction.

It is assumed that the protein is rigid, while the ligand is a small flexible molecule. A fixed coordinate system P is attached to the protein and a moving coordinate system L is defined using three bonded atoms in the ligand as shown in Figure 3.1.1 below. The position and orientation of L relative to P and the torsional angles (i.e. internal degrees of freedom (DOF)) of the ligand define a conformation of the ligand.

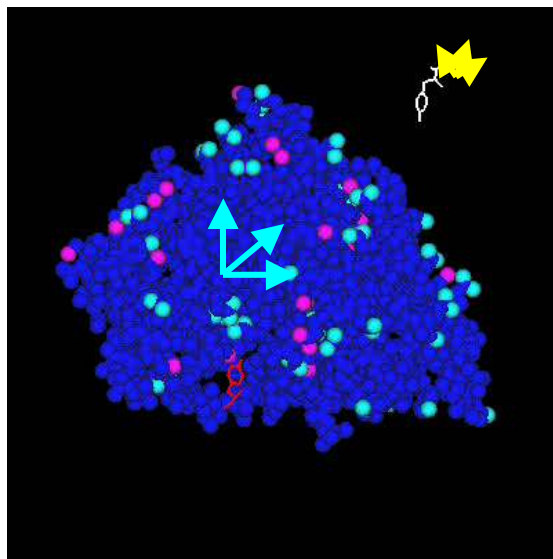


Figure 3.1.1. Fixed coordinate system, P and moving coordinate system, L.

3.1.1. Roadmap Construction (Node Generation)

The nodes of the roadmap are generated by sampling conformations of the ligand uniformly at random, but only in the parameter space around the protein.

The energy E at each sampled conformation is given by the formula below:

$$E = E_{interaction} + E_{internal}$$

where $E_{interaction}$ and $E_{internal}$ are given by the formulas below.

$$\begin{aligned} E_{interaction} &= \text{electrostatic} + \text{van der Waals potential} \\ E_{internal} &= \sum_{\text{non-bonded pairs of atoms}} (\text{electrostatic} + \text{van der Waals potential}) \end{aligned}$$

A sampled conformation is retained as a node in the roadmap with probability, P , given by the expression below. This probability distribution guarantees that there would be a denser distribution of nodes in low-energy regions of the conformation space.

$$P = \begin{cases} 0 & \text{if } E > E_{max} \\ \frac{E - E_{min}}{E_{max} - E_{min}} & \text{if } E_{min} \leq E \leq E_{max} \\ 1 & \text{if } E < E_{min} \end{cases}$$

3.1.2. Roadmap Construction (Edge Generation)

Each node in the roadmap (chosen in the manner described above) is connected to its closest neighbors by straight edges. Each edge is discretized in such a way that between two discrete points on the edge, say q_i and q_{i+1} , no atom moves by more than some threshold, ϵ (usually $1\text{\AA}$.) If the energy at any of the discrete points on the edge is greater than E_{max} , the edge is rejected. For each accepted edge, two weights are assigned. One weight for traversing the edge in one direction and the other for traversing the edge in the opposite direction. Let q be one end of some edge, e , and q' be the other end of e . Then the weight, $w(q \rightarrow q')$ of going from q to q' is given by:

$$w(q \rightarrow q') = \sum_i -\ln(P[q_i \rightarrow q_{i+1}])$$

$$e^{-(E_{i+1} - E_i)/kT}$$

$$\text{where } P[q_i \rightarrow q_{i+1}] =$$

$$\frac{e^{-(E_{i+1} - E_i)/kT}}{e^{-(E_{i+1} - E_i)/kT} + e^{-(E_{i-1} - E_i)/kT}}$$

$P[q_i \rightarrow q_{i+1}]$ is the probability that the ligand will move from q_i to q_{i+1} when it is constrained to move along the edge. $w(q \rightarrow q')$ as defined is a heuristic measure of the energetic difficulty of moving from C to q' . This measure is important for detecting active binding sites, because it captures the difficulty of getting into a binding site, say, $w(q \rightarrow q')$ and the difficulty of leaving it, $w(q' \rightarrow q)$.

3.1.3. Querying the roadmap

For a given goal node q_g like a binding conformation, the Dijkstra's single source algorithm can be used to compute the least-cost path from q_g to each node with a runtime complexity of $O(N \log N)$. N refers to the number of nodes in the roadmap. The least-cost path in the opposite direction can be computed using the same algorithm in the reverse direction.

Other quantities of interest can then be derived from these computations in $O(N)$ time. For instance, the average weights of all paths entering q_g and the average weights of all paths leaving q_g . These averages are akin to the binding and dissociation rates K_{on} and K_{off} .

3.1.4. Experiments and Results

Experiments were run on three complexes namely:

- PDB ID: 1ldm
Receptor: Lactate Dehydrogenase (2386 atoms, 309 residues)
Ligand: Oxamate (6 atoms, 7 dofs)
- PDB ID: 4ts1
Receptor: Mutant of tyrosyl-transfer-RNA synthetase (2423 atoms, 319 residues)
Ligand: L- leucyl-hydroxylamine (13 atoms, 9 dofs)
- PDB ID: 1stp
Receptor: Streptavidin (901 atoms, 121 residues)
Ligand: Biotin (16 atoms, 11 dofs)

As described earlier in section 3.1.1, several thousands of the ligand's conformations around the protein were sampled. Then lowest energy conformations close to the protein surface are selected. Resampling is done around these selections and this process is repeated several times. A number, (approximately 10) of lowest energy conformations whose centers of mass are at least 5Å apart are retained. The table below shows the results from the experiments on 1ldm.

Row number	RMSD from active site configuration (Å)	Configuration energy (kcal/mol)	Avg path weight entering configuration	Avg path weight leaving configuration
0	0.00	-11.79	112.98	134.54
1	31.04	-13.65	85.07	109.94
2	27.49	-12.66	90.48	111.98
3	1.73	-11.72	113.81	137.28
4	28.99	-11.54	85.32	105.19
5	24.67	-11.31	86.26	103.95
6	29.84	-11.27	86.49	107.53
7	29.32	-11.04	85.24	104.64
8	27.07	-10.96	81.70	102.28
9	31.00	-10.13	87.69	104.50
10	28.24	-9.97	86.36	98.89

Table 3.1.4.1. Results from the experiments on 1ldm.

These results show that energy is not a discriminating factor, because some potential sites have lower energy than the active site (first row in the Table 3.1.4.1.) Observe, however, that the average path weights (energetic difficulty) to enter and leave a binding site are significantly greater for the active site. This is an indication that an energy barrier that “traps” the ligand surrounds the active site, i.e. prevents the ligand from leaving the active site.

3.2. Application of Roadmaps to Protein Folding

This was work done by N.M. Amato, K.A. Dill, and G. Song and can be used to find the order of SSE formation in a protein. Determining the order of formation of secondary structures is done through a 3-step process, the mechanics of which are described in section 3.2.2. A high-level overview the process is as follows: first, the least-cost path from each denatured conformation to the folded one is extracted. These costs are the weights associated with the edges in the roadmap. Then, the order of formation of SSE's is computed along each path. Finally, the formation order that appears most often over all paths is assumed to be the SSE formation order of the protein.

Note that the process described above returns a partial ordering (as opposed to a total ordering), since there is uncertainty as to when a particular structure appears due to variance.

First, a known native state is assumed and the kinematic model used here is based on degrees of freedom of the protein, i.e. the ϕ - ψ angles. The energy function consists of van der waals, hydrogen bonds and hydrophobic effect terms. The sampling strategy here is different and will be discussed next. Edge generation remains the same as described in section 3.1.2.

3.2.1. Roadmap Construction (Node Generation)

The sampling space here is non-uniform, using a gaussian distribution instead to sample the conformations around the known native state. The conformations are sorted into bins based on the number of native contacts. Native contacts are defined as pairs of C α atoms that are close together in the native structure. The sampling process ends when all the bins have some minimum number of conformations. This implies a “good” coverage of the sample space.

3.2.2. Querying the roadmap

A contact map showing the time step when each native contact appears is built. This time step is approximated as the average of the appearance time steps of its contacts, each edge is a time step. Clearly, this means that outliers could unduly influence these time steps. Also, traversing some edges in the roadmap might take more wall clock time than others. Nevertheless, the results from the experiments agree with the findings using NMR spectrometers.

3.2.3. Experiments and Results

Secondary Structure Formation Order and Validation					
pdb	res. #	SS	secondary structure formation order	roadmap	exp.[25]
1GB1	56	1 α +	α 1, β 3- β 4, β 1- β 2, β 1- β 4	5126	Agreed
1BDD	60	3	α 2, α 3, α 1, α 2- α 3, α 1- α 3	5471	Agreed
1COA	64	1 α +	α 1, β 3- β 4, β 2- β 3, β 1- β 4, α 1- β 4	7975	Agreed
1UBQ	76	1 α +	α 1, β 3- β 4, β 1- β 2, β 3- β 5, β 1- β 5	8357	Agreed

Table 3.2.3.1. Results showing that the computation method agrees with NMR results.

The results above show that the predictions of the roadmap are in agreement with the results generated by the NMR spectrometer.

3.3. Stochastic Roadmaps

This was work done by M.S. Apaydin, D.L. Brutlag, C. Guestrin, D. Hsu, J.C. Latombe, and C. Varma. The new idea here was to capture the stochastic nature of molecular motion by assigning probabilities to edges. These edge probabilities, P_{ij} , follow the metropolis criteria and are defined as follows:

$$P_{ij} = \begin{cases} \frac{\exp(-\Delta E_{ij} / kT)}{|N_i|} & \text{if } \Delta E_{ij} > 0 \\ \frac{1}{|N_i|} & \text{otherwise} \end{cases}$$

Self-transition probability, P_{ii} is defined as: $P_{ii} = 1 - \sum_{j \neq i} P_{ij}$

In this formulation, roadmap nodes are sampled uniformly at random like method 1 of section 3.1, but the energy profile along the edges is not considered unlike method 1.

3.3.1. Stochastic Roadmap Simulation (SRS)

Stochastic roadmap simulation and Monte Carlo simulation both converge to the Boltzmann distribution. That is, the number of times SRS is at a node in a set of nodes, V , converges toward

$\int_V e^{-E/kT} dV$ when the number of nodes grows and they are uniformly distributed. However, SRS is a lot faster than MC simulation, since the energy model used by SRS is relatively simple and MC has to do many energy computations to escape local minima.

The transition probabilities described earlier depend only on i and j , i.e. they are only dependent on the two nodes involved. This means that the roadmap defines a Markov chain. There is one linear equation per node, and the solution to the linear system gives the P_{fold} for all nodes. The linear system thus defined is sparse, since each node is only connected to some constant number of nodes.

Note that no explicit simulation is run, but all pathways are taken into account. The roadmap encodes multiple different pathways and can be seen from the example drawn from <http://mathworld.wolfram.com/Self-AvoidingWalk.html> on the number of self-avoiding walk on a 2-D grid.

3.3.2. Experiments and Results

One of the results from experiments done on a β hairpin is shown below. Others can be seen on the lecture slides.

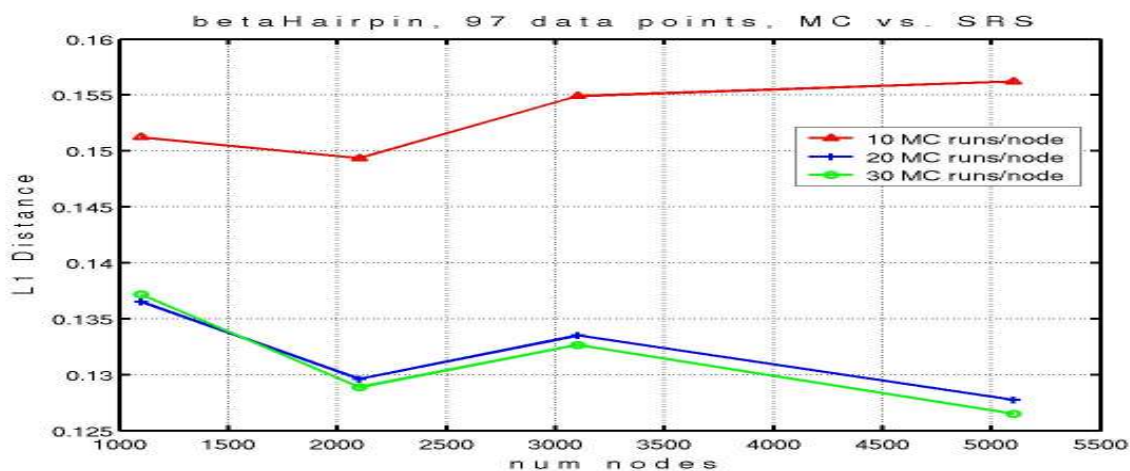


Figure 3.3.2.1. Results from approximately 100 conformations of the β hairpin.

	# of conformations	Computer time taken	# of energy computations
Monte Carlo (30 Simulations)	1	10 hours	Over 10, 000, 000
Roadmap	2000	23 seconds	Approx. 50,000

Table 3.3.2.1. Computation time for approximately 100 conformations of the β hairpin.

These results show a speedup factor of approximately 6 going from MC simulations to SRS!

3.4. Using Path Sampling to Construct Roadmaps

This work was done by N. Singhal, C.D. Snow, and V.S. Pande. Here, paths computed with molecular dynamics simulation techniques are used to create the nodes of the roadmap. This process is called path shooting. One example of an MD path would be one generated by Langevin dynamics equation of motion. The intuition is that path shooting should yield more pertinent / better distributed nodes. Edges are labeled with the time needed to traverse them.

If two nodes are closer than some threshold, ϵ , they are merged into one and merging rules are applied to update edge probabilities and times. This creates an approximate distribution of nodes over the reachable subset of conformational space.

This has applications in the computation of Mean First Passage Time (MFPT). This is the average time when a protein first reaches its folded state.

3.4.1. Experiments and Results

Using [Folding@Home](#) (a cluster of computers used for MC simulations) on 12-residue tryptophan zipper beta hairpin (TZ2), 1750 trajectories (fully atomistic simulation) ranging from 10 to 450 ns were generated. Of these trajectories, 14 reaching folded state. The resulting roadmap had 22,400 nodes. MFPT was found to be approximately 2-9 ms, which is similar to experimental measurements (from fluorescence and IR.) These results were the first set of quantitative results and are certainly more satisfying.

3.5. Conformational Analysis of Long Protein Loops

This work was done by J. Cortés, T. Siméon, M. Renaud-Siméon, and V. Tran. Tree-shaped roadmaps are built to explore the clash-free subset of the conformational space of a loop. The kinematic model is made up of the ϕ - ψ angles on the backbone and the χ_i torsional angles in side chains. The enzyme used was Amylosucrase (AS), which is the only enzyme in its family that acts on sucrose substrate. The 17-residue loop (named loop 7) between Gly433 and Gly449 is believed to play a pivotal role. Amylosucrase is shown below.

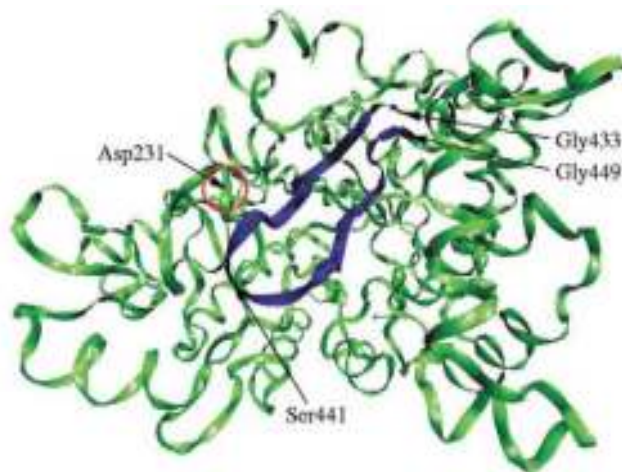


Figure 3.5.1. Amylosucrase with loop 7 shown in blue.

3.5.1. Roadmap Construction

A tree-shaped roadmap is created from a start conformation q_{start} . At each step of the roadmap construction, a conformation q_{rand} of the loop is picked at random and a new roadmap node is created by iteratively pulling it toward the existing node that is closest to q_{rand} . The iteration ends when it becomes impossible to get closer to q_{rand} or a clash is detected.

3.5.2. Experiments and Results

Surprisingly, loop 7 cannot move much. Main bottleneck is residue Asp231. If residue Asp231 is removed, then loop 7's mobility increases dramatically. The C α atom of Ser441 can be displaced by more than 9Å from its crystallographic position.

4. Conclusion

Probabilistic roadmaps are a recent, but promising tool for exploring conformational space and computing ensemble properties of molecular pathways. Some ideas that are currently being explored, or should be explored in future include:

- Better sampling strategies able to handle more complex molecular models (protein-protein binding).
- More work to include time information in roadmaps, instead of using the number of edges to approximate the amount of time it takes to traverse a path.
- More thorough experimental validation to compare computed and measured quantitative properties

References

1. Du, Pande, Grosberg, Tanaka, and Shakhnovich. On the Transition Coordinate for Protein Folding. Journal of Chemical Physics (1998).