Capturing Molecular Energy Landscapes with Probabilistic Conformational Roadmaps

**Mehmet Serkan Apaydin[1], Amit P. Singh[2], Douglas L. Brutlag[2],
and Jean-Claude Latombe[1]**

*Departments of Computer Science[1] and Biochemistry[2]
Stanford University, Stanford, CA 94305, USA*

Abstract: Probabilistic roadmaps are an effective tool to compute the connectivity of the collision-free subset of high-dimensional robot configuration space. This paper extends them to capture the pertinent features of continuous functions over high-dimensional spaces. This extension has several possible applications, but the focus here is on computing energetically favorable motions of bio-molecules. Many bio-chemical processes essential to life require certain molecules to adopt different shapes over time. Computational tools predicting such motions can help better understand these processes and design useful molecules (e.g., new drugs). In this context, a molecule is modeled as an articulated structure moving in an energy field. The set of all its 3-D placements is the molecule's conformational space, over which the energy field is defined. A probabilistic conformational roadmap (PCR) tries to capture the connectivity of the low-energy subset of a conformational space, in the form of a network of weighted local pathways. The weight of a pathway measures the energetic difficulty for the molecule to move along it. The power of a PCR derives from its ability to compactly encode a large number of energetically favorable molecular pathways, each defined as a sequence of contiguous local pathways. This paper describes general techniques to compute and query PCRs, and presents implementations to study ligand-protein binding and protein folding.

1. Introduction and Motivation

An insight from research in biology is that the function of a bio-molecule follows from its form. For instance, to act as a potent inhibitor, a drug molecule must bind solidly against a protein's cavity (the binding site), which requires that the molecular surfaces in contact have close steric and coulombic match [SK93]. In addition, molecules are neither static, nor rigid. In fact, chemical processes essential to life depend on the ability of certain molecules to adopt different shapes over time. E.g., a drug molecule must both move to eventually reach a binding site and deform to achieve a conformation that fits well and lock into this site (docking process). Molecular movements occur under the influence of forces induced by energy fields.[Hai92].

Computational models able to effectively simulate and predict molecular motions have important potential applications, notably in drug and protein design. For instance, being able to reliably simulate the ligand-protein docking process would make it possible to automatically extract promising drug candidates (leads) from large existing databases of ligands [LFKL00] and test the docking abilities of variants of these leads.

Molecules can be modeled as articulated structures made of spheres (representing atoms) connected by links (bonds between atoms). The main degrees of freedom (dofs) are torsional dofs about some links. Consider three consecutive links v_1, v_2, and v_3. The torsional dof around v_2 corresponds to varying the dihedral angle made by the plane containing v_1 and v_2 and the plane containing v_2 and v_3. The assignment of an angular value to each dof defines a *conformation* of the molecule (a concept similar to that of a configuration in robotics). The set of all conformations is the molecule's conformational space, which has as many dimensions as there are dofs.

While drug molecules typically consist of 10 to 50 atoms, with 5 to 15 torsional dofs, proteins contain thousands to hundreds of thousands of atoms, with hundreds to thousands of dofs. Energy fields are defined over the conformational spaces of the molecules (or the cross product of several such spaces, if we consider multiple molecules interacting and deforming simultaneously). Molecular movements are described as continuous pathways in a conformational space.

Many computational models assume that molecules are rigid structures. Such models may tell us that a ligand fits into a protein's active site, but give no indication of the conformational changes of the ligand and/or the protein that were required to achieve their final bound state. In reality, the docking may not even be possible because it would require a molecule to traverse high-energy conformations. The transition from static to dynamic models brings us into molecular dynamics [Hai92]. In theory, the energy fields causing molecular motions are well understood. But the precise simulation of these motions over the time periods during which the phenomena of interest take place is well beyond the capabilities of today's fastest computers [DK98]. Indeed, many dofs often participate in molecular movements, while energy functions include a huge number of terms usually involving combinations between all pairs of atoms. To ensure simulation accuracy, time steps taken by molecular dynamics techniques are usually on the order of femtoseconds. Taking the solvent into account further adds to this complexity.

Researchers have developed approximate energy models that are less expensive to compute, e.g., by using principal component analysis to detect important dofs and "freeze" the others [TPK00], ignoring energy terms involving atoms that are some distance apart, and/or treating groups of atoms (e.g., rings, side-chains, α-helices) as single units. Simulation techniques connecting local minima of such models produce plausible pathways at a more reasonable cost. Some randomness may be introduced into the computation to account for model imperfection [KBK94].

Classical simulation techniques lead to generating individual pathways such as the one shown in Figure 1. However, the number of pathways that can be computed in practice is rather small. Instead, our goal is to capture the relevant characteristics of the "energy landscape" over the conformational space by a network of pathways. This network, called a *probabilistic conformational roadmap* (PCR), is a graph whose nodes and edges are respectively low-energy conformations and short
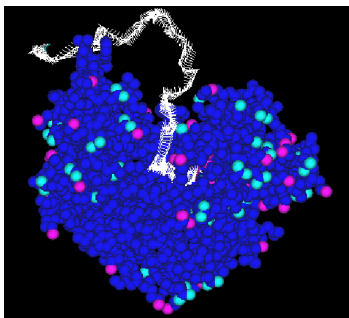
Figure 1. Ligand docking against a protein

weighted pathways. The weight of a pathway estimates the energetic difficulty for the molecule(s) to move between the two conformations. By combining a large number of short paths, a PCR compactly encodes a large number of energetically plausible paths. It thus has the ability to represent the pertinent energy landscape over the conformational space in a form that is more directly exploitable than the original energy function. Once computed, a PCR may be queried in a variety of ways.

Probabilistic techniques (combined with optimization and clustering) have been used to sample conformational spaces of ligands and identify their low-energy conformations [FHK96]. But they do not attempt to connect the sampled conformations into a network. PCRs are a rather direct extension of the probabilistic roadmaps introduced in [KSLO96]. However, while robot configurations are admissible (e.g., collision-free), or not, a conformational space is the domain of a continuous energy function, where lower-energy conformations are more favorable. Therefore, while a classical probabilistic roadmap tries to capture the landscape of a binary function, a PCR has a similar goal, but with a more complex function. The method in [KB99], which constructs a roadmap connecting local minima of a potential function, has some resemblance with ours. But it connects local minima using an up-hill search technique to climb out of local minima towards saddle points. This operation, which requires many evaluations of the potential function, would be too expensive in our case. The concept of a PCR was first introduced in [SLB99], along with its application to ligand-protein binding. The present paper extends this concept, provides new results for ligand-protein binding, and explores the application of PCRs to protein folding. Other ongoing research aimed at applying PCRs to ligand-protein binding and protein folding is reported in [BSA00, SA00].

The problem of capturing functional landscapes over complex spaces is one of general interest. For example, outdoor mobile robots must compute motion plans that take the navigability of the local terrain into account (e.g., muddy and steep areas are more difficult to traverse than flat hard terrain). The navigability of a terrain often depends on the heading of the robot and is best defined over the robot's configuration space. Algorithms have been proposed to compute paths with acceptable or optimal characteristics [MM97]. Roadmaps similar to PCRs could better capture the pertinent properties of the navigability function over the configuration spaces. Another application is minimally-invasive surgical planning, where one must plan the paths of surgical instruments (e.g., scalpels, endoscopes) to minimize damage on healthy tissue, with some tissues (e.g., blood vessels) being more critical than others (e.g., fat).

Section 2 outlines the basic PCR framework. Sections 3 and 4 apply this framework to two problems, ligand-protein binding and protein folding.

2. Probabilistic Conformational Roadmap

2.1. Classical Probabilistic Roadmap
A classical probabilistic roadmap R is created over a robot's configuration space C [KSLO96]. R is a graph whose nodes are points of C (called *milestones*) and edges are short simple paths (*local paths*) between milestones. The local paths are usually straight-line segments. Points in C are either admissible (e.g., collision-free), or non-admissible. R should lie in the admissible subset C_A of C and capture the connectivity of C_A as well as possible. Ideally, there should be a one-to-one correspondence between the connected components of R and those of C_A, and every point in C_A should be connectable to a milestone by a simple path [HLMK99].

The roadmap R is computed as follows. The range of each dof parameter is normalized so that $C = [0,1]^n$, where n is the number of dofs. Points are picked at random in $[0,1]^n$ and the admissible ones are retained as milestones. Next, pairs of milestones that are sufficiently close to one another are considered and for each pair a local path connecting the two milestones is tested for admissibility. If this path lies in C_A, an edge of R is created between the two corresponding milestones. This basic scheme admits many variants. Points may be picked from $[0,1]^n$ uniformly, or using more sophisticated probabilistic distributions.

Theoretical analysis shows that under reasonable assumptions the probability that a probabilistic roadmap made of s milestones fails to correctly capture the connectivity of a given space C_A converges toward 0 as e^{-s}. In practice, probabilistic roadmaps have been used successfully to solve motion-planning problems in high-dimensional spaces and/or in the presence of complex admissibility constraints.

2.2. Probabilistic Conformational Roadmaps
Let C be the conformational space of a molecule or a group of interacting molecules. If we study protein folding, C may be the conformational space of the protein of interest. But for practical reasons, C may only encode a subset of the protein's dofs, e.g., by considering every amino-acid side-chain as a rigid unit. If we study ligand-protein binding, C may encode dofs of both the ligand and the protein, or it may only be the ligand's conformational space if we assume that the protein does not deform significantly during the docking process.

Let $E: C \to \mathbf{R}$ be a potential energy field over C. E may combine terms that express a molecule's own potential energy and terms that relate to the interaction between molecules To illustrate, Figure 2a shows an imaginary function E over a two-dimensional space $C = [0,1]^2$. E varies between -47 and +52.5.

We need a metric over C, such as the maximal distance between two corresponding atoms. But others would do as well.

A PCR is constructed by picking points from C uniformly at random. This is done by assigning random values to each coordinate of C, within its given range of possible values. For each point q we compute $E(q)$ and we accept q as a milestone of the PCR at random with the following probability distribution:

- 0 if $E(q) > E_{max}$
- $(E_{max} - E(q))/(E_{max} - E_{min})$ if $E_{max} \geq E(q) \geq E_{min}$
- 1 if $E_{min} > E(q)$

The resulting milestone distribution is denser in low-energy regions of C.

Let s be the number of milestones selected as above. The next step is to connect every milestone by local paths to at most k other milestones, where k is selected roughly equal to the number of dimensions of C, so that the resulting PCR has size linear in s. The connection algorithm is the following:

For $i = 1, 2, \ldots, s\text{-}1$

1. Set Q to be the queue of the K milestones m_j ($j > i$) that are closest to m_i, sorted according to their distance to m_i

2. While the number of edges at m_i is less than k and Q is not empty
   a. $m \leftarrow \text{extract}(Q)$
   b. If the straight-line segment (local path) between m and m_i lies in a low-energy region, then connect m and m_i by an edge

We implement Step 2.b by discretizing the segment into a series of points spaced by some small distance ε. An edge is generated if all these points have energy less than a given threshold. Hence, local paths that traverse a high-energy barrier are discarded. Step 2.b is potentially expensive, as it requires computing the energy function at multiple points. So, we bound the number of times it is executed by K for each milestone. K is set to 3 to 5 times k.

Figure 2b shows the projection of a PCR computed by this algorithm for the landscape of Figure 2a, with $k = 4$, $K = 12$, $E_{min} = -40$ and $E_{max} = 20$. The metric used here is the Euclidean distance in $\mathbf{R}^2$.

Finally, for every pair of connected milestones m and m', we estimate the likelihood of the molecule(s) to transit along the local path t joining them. Let p be the number of discretized points generated at resolution ε along t and $E_1, \ldots, E_p$ be the values of E already computed at these points. For any three successive points q_{i-1}, q_i, and q_{i+1} we use the following equation to estimate the probability of moving from q_i to q_{i+1}:

$$\Pr[q_i, q_{i+1}] = \frac{e^{-(E_{i+1} - E_i)/kT}}{e^{-(E_{i+1} - E_i)/kT} + e^{-(E_{i-1} - E_i)/kT}} .$$

We compute the weight of t as: $w = -\Sigma_{i=1 \text{ to } p} \log(\Pr[q_i, q_{i+1}])$.

Local paths through higher-energy conformations have higher weights than those lying entirely in a low-energy area. A term proportional to the path length is added to w. As the total weight is not the same in both directions. We compute and store both weights. During query processing, we select the weight corresponding to the direction in which the path is traversed.

The construction of a PCR requires choosing several parameters: s, k, K, E_{min}, E_{max}, and ε. The most difficult to select is the number s of milestones. We do not know how big should s be for the PCR to effectively capture the landscape of C. An exponential rate of convergence has been formally established for classical probabilistic roadmaps [HLMK99], but no such result has been proven yet for PCRs.
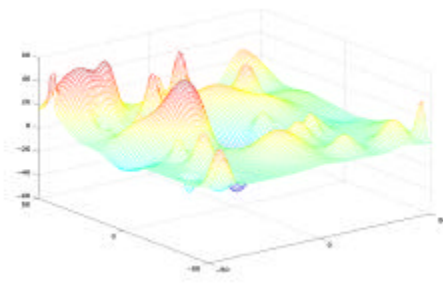
(a)
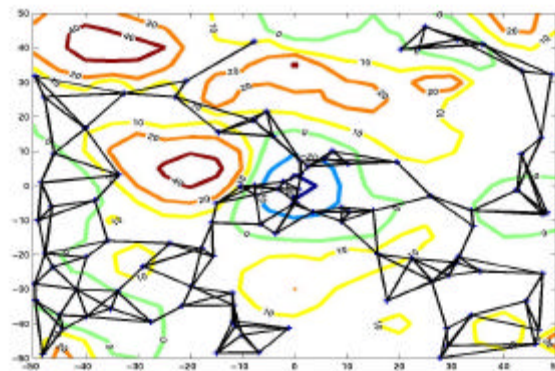
(b)

Figure 2: Fictitious energy function and computed PCR

2.3. Querying a PCR

It is important to note that the information contained in R cannot be better than the energy function E used to construct it. Furthermore, one could obtain individual pathways reflecting the energy function more precisely, by tracking the values of E at a fine resolution in the conformational space, rather than by sequencing local pathways contained in the PCR. The main advantage of a PCR is that it encodes a large number of paths scattered across the energetically favorable regions of C. Hence, though E is imperfect and milestones provide relatively low-resolution sampling of C, one may get significant and reliable information by collecting invariants and statistics from subsets of paths contained in the PCR. Query processing should take advantage of this strength, instead of relying on individual paths .

The most straightforward query is to determine if energetically favorable paths exist between two input conformations. Defining the weight of a pathway to be the sum of the weights of the local paths it contains, a search algorithm finds N best paths in the PCR between the two input conformations (for some given N). These paths can be visualized and statistics can be computed (e.g., number of milestones, average weight, energy profile). To avoid showing many similar results, the paths can be grouped into clusters using a similarity metric and only one path in each cluster may be output. If the same milestone m lies on several such paths, m may be considered as a likely intermediate, hence a potentially relevant biological structure.

Another query is to find N best paths that enter an input goal conformation, and display the milestones that are contained in those paths. Again, similar paths can be grouped into clusters, and milestones that lie on several distinct paths can be identified as likely intermediates. The average weight of these paths can be compared to the average path weight for other possible end

3

conformations, in order to provide insight on why an end conformation is more likely to be attained than another.

One may use a PCR to perform stochastic simulations. Starting at some initial conformation, a simulation run proceeds step by step. At each step, it decides at random to either stay at the current milestone or transit to an adjacent, using a probability distribution based on the weights of the local paths. This type of simulation corresponds well to the modeled biological process since the molecules do not have the prior knowledge of their final conformations. Multiple simulation runs can be performed and statistics can be collected about the traversed milestones. Further analysis may also help discover funnels of attraction steering a molecule toward an end conformation.

2.4. Computational Enhancements

The cost of computing a PCR dominates that of performing many queries. It is therefore desirable to develop techniques that can produce good PCRs faster. One technique, which was successfully applied to classical roadmap [KSLO96], is to construct a roadmap in two stages. A first roadmap R_1 is created with $s_1 < s$ milestones. Then, $s_2 = s - s_1$ milestones (and the corresponding connections) are added to R_1 to form the final roadmap R. The new milestones are picked at random around milestones of R_1 that are the least connected to other milestones (e.g., the number of connections is less than k). Another technique is to evaluate the energy function at a few conformations around every low-energy milestone m and, if important energy variations are detected, to pick new milestones around m. This may help build denser PCRs around important states.

One may use multiple energy models of different complexity. Suppose that a function E' is available, which approximates the energy function E, but at a fraction of the computational cost of E. Let $H(E)$ and $H(E')$ designate the respective subsets of C over which E and E' take high values. We would like $H(E') \subseteq H(E)$ and the difference between the two subsets to be rather small. Since a molecular energy function such as E contains many terms, it is often possible to build E' by ignoring a large number of relatively small terms. We can use E' to build a roadmap R' of size s' much greater than s. Next, we re-consider the milestones and connections in R' and accept/reject them using E to produce the final roadmap R. If E' costs one or two orders of magnitude less time to evaluate than E, we can obtain a PCR of given size s in much less time than by generating it using the only function E.

Each time a point is picked at random in C, a gradient technique could track the steepest descent of E and generate a point of lower energy. This new point would then be the actual milestone candidate. Obviously, the cost of generating each milestone would be greater, and this cost would have to be weighted against that of generating more milestones (without local optimization). Local paths could also be improved by iteratively deforming them into curved ones in order to minimize their weights. Other improvements may use prior knowledge about the molecules and/or the molecular process. For instance, atomic symmetries that cause some torsional angles to have preferred values may be detected. Milestones can then be generated by selecting these angles using non-uniform distributions with peaks at the preferred values. If one knows in advance critical low-energy conformations, such as a ligand's binding conformation or a protein's folded state, these conformations can be input as milestones. A greater density of additional milestones may be generated around them since they often lie in convoluted low-energy passageways [LGH97], which may be difficult to capture by a standard sampling technique [HKL98].

3. Ligand-Protein Binding

3.1. Problem Statement

Biomolecular interactions, such as molecular binding, are critical to the process of life. Ligand-protein binding involves a small molecule (10-100 atoms) – the ligand -- binding to a specific site on a larger receptor protein. Ligands are used for signaling and regulation in virtually all cellular pathways. Most drug molecules are ligands that inhibit or enhance the activity taking place at the protein sites where they bind. For instance, it was discovered that a specific enzyme (protein) -- the HIV-1 protease -- cleaves the amino-acid chains produced by the HIV virus, hence playing an essential role in the life cycle of this virus. Drugs have been designed which bind to the active site of the HIV-1 protease and thus physically block the amino-acid chains produced by the HIV virus from entering this site.

Most techniques to predict ligand-protein binding attempt to compute the final conformation of the ligand by maximizing an energy score and do not explicitly study the dynamic or kinetic properties of the binding process. To study such properties, researchers have relied on Molecular dynamics, Brownian Dynamics and Monte Carlo simulation techniques. However, these techniques are computationally intensive, especially for ligands with many dofs, and provide, at best, a small number of plausible ligand's pathways.

3.2. Application of PCR

PCRs offer a novel approach to studying the dynamics and kinetics of the ligand-protein docking process by sampling from the space of *all* possible paths that a ligand may follow as it binds to the receptor protein. Hence, instead of simulating the docking process, we use a PCR to effectively guess several possible intermediate conformations of the ligand and obtain a distribution of energetically favorable paths to the binding site via these intermediate conformations.

In the following, we assume that the protein is rigid. This assumption allows us to generate PCRs in the ligand's conformation space. There are cases where deformations of the protein could not be ignored [TPK00]. In those cases, one must consider a conformation space encoding both the ligand's dofs and some protein's dofs.

We model the ligand as an articulated linkage made of spheres (atoms) connected by straight links (bonds). An arbitrarily chosen terminal atom is given 5 dofs, 3 specifying its center's coordinates and 2 specifying the orientation of its only bond. Each other dof is a torsional dof around a bond between two non-terminal atoms. As angles between two successive bonds and bond lengths usually undergo very small variations, we assume they are constant. Atomic rings are modeled as rigid units, which is true in most organic molecules. Terminal hydrogen atoms are not explicitly modeled, but are accounted for by increasing the radius of the atoms they are bonded to. The 5 dofs of the root atom and the torsional dofs define the conformational space C.

Our energy model over *C* consists of two components: the energy of interaction of the ligand with the receptor and the internal energy of the ligand. For a given point in *C*, the energy of interaction is computed by first calculating the coordinates of the ligand atoms in a fixed coordinate system. The energy contributions of each ligand atom are computed based on the potential field created by the protein at the atom's coordinates. This field is calculated from the atom coordinates and charge distribution of the protein. It consists of the van der Waals potential and the electrostatic potential. The van der Waals potential represents the steric constraints on atomic interactions and is modeled using the following Lennard-Jones 12-6 function:

$$v(r) = e\left\{\left(\frac{r_0}{r}\right)^{12} - 2\left(\frac{r_0}{r}\right)^6\right\},$$

where r is the distance between two atoms, r_0 is the distance at which the energy is minimum, and ε is the well depth, i.e., $-\mathbf{n}(r_0)$, usually about 0.2 kcal/mol.

Since the standard Coulombic equation of electrostatic interaction is valid only for an infinite medium of uniform dielectric, it cannot be used here. The dielectric discontinuity between protein and solvent generates induced or reflected charges that can play a significant role in the binding process. Hence, we model electrostatics using the following Poisson-Boltzmann equation, which is a widely accepted model of electrostatic interactions in solution:

$$\nabla \cdot [e(r) \nabla \cdot \ddot{o}(r)] - e(r)\mathbf{k}(r)^2 \sinh[\mathbf{f}(r)] + 4\mathbf{p} r^f(r)/kT = 0$$

where $\mathbf{f}$ is the electrostatic potential in units of kT/q, k is the Boltzmann constant, T is the absolute temperature, q is the charge on a proton, ε is the dielectric constant, and ρ^f is the fixed charge density. We use the Delphi program [SH90] to solve the equation on a 3-D grid at a resolution of 0.5Å. The van der Waals potentials are computed at the same grid resolution by calculating for each grid point the potential contribution of all receptor atoms within a threshold distance of 10Å.

We compute the energy of interaction of every ligand atom with the protein by indexing the atom's center to the nearest grid point and retrieving the van der Waals and electrostatic potentials at this point. The total energy of interaction is computed by summing the contributions of each atom. The ligand's internal energy is computed by applying the standard van der Waals and Coulombic equations to each non-bonded pair of ligand atoms. (Since a ligand is small and flexible, we assume that its surface is not well defined and hence use the standard Coulombic equation, with a dielectric constant between 60-80.)

Milestones are generated as described in Section 2.2. In addition, extra milestones are generated by iteratively over-sampling regions of lowest energy in *C*. The final bound state of the ligand is also entered into the roadmap as a milestone.

3.3. Experimental Results

We have constructed PCRs for various ligand-protein complexes. Initial tests were performed on three complexes identified as 1ldm, 4ts1, and 1stp in the Protein Data Bank (PDB, http://www.rcsb.org). Further tests were carried out on complexes that appear to be mediated primarily by electrostatic effects (e.g., superoxide dismutase and acetylcholine esterase). The PCRs were constructed with 5,000 to 100,000 milestones. On a 195-MHz MIPS R10000 processor, the average PCR construction
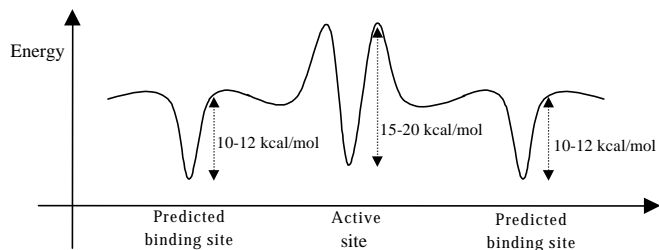

Figure 3: Illustration of energy barriers around an active site

time ranged from 3-8 minutes for smaller roadmaps to 0.5-3 hours for larger ones. In all runs, our software generated PCRs containing 2 to 5 connected components. As voids and narrow cavities do occur within a protein structure or on its surface, some milestones may be picked in these regions, thus yielding more than one connected component. In each run, over 98% of the milestones were in a single component also containing the bound state.

We studied if our software was able to distinguish the active site from other low-energy sites. The two attributes we used to distinguish between the active and other predicted binding conformations were the ligand's absolute energy and the average weight of the paths entering and leaving the conformation. The average weight was computed by generating many paths from randomly selected conformations to the final conformation.

We observed that the absolute energy of the ligand was not a strong discriminating factor between the active site and other predicted sites. In two of our three test cases (1ldm and 1stp) the algorithm found ligand conformations outside the active site with energies equal to or even slightly lower than the ligand's energy in its active site conformation. Instead, using the average path weight, our software was able to distinguish between the active site and other predicted sites. The average weight of all paths entering and leaving the active site was on average 30% higher than the weights for all other low-energy sites. Therefore, while it is significantly more difficult for the ligand to *leave* the active site than the other low-energy binding sites, it is also more difficult for the ligand to *enter* the active site. We believe that this result indicates the presence of an energy barrier around the active site that traps the ligand within the site. Figure 3 shows a schematic of a possible energy contour that could yield a similar result. Our experiments also show that the average weight of paths entering the active site is of the same order as the weight of paths leaving the predicted sites (a result reflected in Figure 3). Hence, the difficulty of entering the active site is approximately equal to the difficulty of leaving the other binding sites.

Other tests have focused on analyzing the role of the electrostatic energy in binding. In one series of experiments, we have selectively eliminated one or more of the charges on the protein. When all charges are turned off, the results show that the energy barrier we previously detected is largely eliminated. Hence, not only do the energy minima in the binding site increase, but the energy barrier surrounding the binding site also seem to decrease, hence flattening the curve in Figure 3. These results indicate that the barriers to ligand docking are mainly of electrostatic nature, and not caused by van der Waals potentials. In addition, we have stochastically simulated the motion of the ligand in a PCR by selecting paths from each milestone based on the distribution of outbound local path weights. Initial results
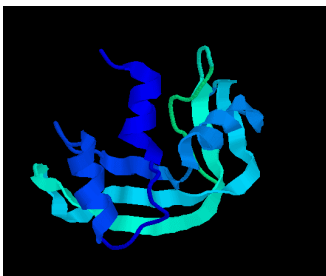
Figure 4. Secondary structure of the ribonuclease A

indicate that electrostatics steering is detectable by the PCR, but only at short distances from the molecular surface (5-7Å).

4. Protein Folding

## 4.1. Problem Statement

A protein is a sequence of amino acids that folds to generate a compact 3-D structure. This structure performs many functions, from building larger assemblies such as muscle fibers to providing specific binding sites for other molecules. The position of the atoms in a folded protein is referred to as the protein's tertiary structure. The primary structure is the amino-acid sequence, while the secondary structure refers to specific local arrangements of a few to a few dozen amino acids. There are two main types of secondary structure elements (SSE): *a-helices* and *b-strands*. These SSEs have regular structures, with repeating torsion angles and a constant pattern of hydrogen bonds. They are usually connected by *loops*, which have irregular shapes. An α-helix has a corkscrew shape, with the atoms on the backbone closely packed and the side-chains extended in a helical array. A β-strand is an almost fully extended series of 5 to 10 amino acids. Two or more β-strands often align side-by-side into a *b-sheet* held together by hydrogen bonds. Most folded proteins are a sequence of α-helices and β-strands connected by loops. Figure 4 shows the secondary structure of the ribonuclease A (a digestive enzyme). Note how intricately and compactly the SSEs are interwoven.

Recent advances in X-ray crystallography and NMR imaging have made it possible to elucidate the folded conformations of a rapidly increasing number of proteins. However, little is known today about the folding pathways that transform an extended string of amino acids into a compact and stable structure. So far it has only been possible to identify approximate intermediate conformations for few proteins. Some biological experiments track a particular property of the protein during folding, but they provide a limited way of following the folding pathway. The ability to predict pathways would help design proteins with desirable properties [KL99]. It could also help determine why relatively small alterations in amino acids may result in dramatic changes of a protein's folded state. Several diseases such as Cretzfeldt-Jakob's, Alzheimer's, and cystic fibrosis are believed to be the result of protein misfolding.

4.2. Application of PCR

The application of PCRs to protein folding is made complex by the large number of dofs. To simplify, we assume here that the SSEs of the protein have already formed and are given as inputs. This assumption loosely corresponds to studying the folding process after the protein has acquired the so-called "molten
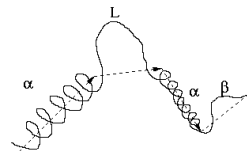
**Figure 5: Representation of a protein**

globule" state, an observed intermediate for some proteins [PR97]. This state has nearly the same secondary structure as the final fold, but the tertiary structure is not as compact. The pathways provided by a PCR based on this assumption could help understand how α-helices and β-sheets interweave into a compact geometric arrangement.

In a similar way to [SB97], we represent the protein as a sequence of vectors, each representing an SSE (Figure 5). We consider the following dofs (Figure 6):

- A *revolute* dof is located at the extremity of each vector, except the last one. The corresponding parameter is the angle made by the vectors ending and starting at this point.
- A *dihedral* dof is associated to every three consecutive vectors. The parameter is the angle made by the plane containing the first two vectors and the plane containing the last two.
- A *twist* dof is associated to every α-helix and β-strand. A coordinate frame is attached to this SSE with its z axis aligned with the element vector. The dof parameter is the angle between the x axis of this frame and the orientation of the first amino acid on that vector. The twist of an α-helix or β-strand about its own axis does not affect the positions and orientations of other SSEs.
- A *prismatic* dof is associated with each loop. The parameter is the length of the loop vector, which is allowed to vary within a range that is a function of the number of amino acids in the loop.

Our potential function is taken from [STD95]. It has a hydrophobic-interaction and an excluded-volume part. Amino acids are categorized into two groups, hydrophobic (H) and hydrophilic (or polar, P). H-H contacts are favorable, whereas H-P or P-P contacts do not contribute to the energy. The exclusion term ensures that no two atoms are too close. There are also terms of a third type for β-sheets, which account for hydrogen bonding. These terms are a function of the distances between side-chain centroids. This model assumes that hydrophobic interactions drive the folding process and that the specific identity of the side-chains is only responsible for the fine-tuning of the fold. It is argued in [STD95] that the level of success of their function is comparable to that of functions with hundreds to thousands of parameters.

We generate each milestone of a PCR by sampling each dof of the protein's model at random. We explicitly input the folded state as a milestone. Lower energy conformations have a higher probability of being accepted. We only compare the exclusion energy component, rather than the total energy, in accepting a conformation. The reason for this is that H-H interactions may counterbalance the contribution of mutually close side-chains to
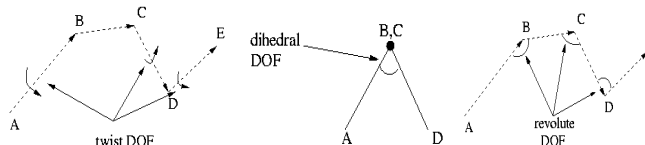
Figure 6: Degrees of freedom in our protein model

the energy, thus resulting in self-colliding conformations. We set the exclusion energy threshold to prevent any two side-chains from coming closer than 3.8Å. We also take extra samples around the given folded state.

Each milestone is then connected to at most k milestones among the nearest ones, where k is the number of dof in our protein's model. The nearest neighbors of a milestone are found using ANN [AM93], with the Euclidean distance over the conformational space, after normalizing each dof parameter to lie between 0 and 1. We also tried the RMSD metric, but it was slower and did not give significantly different results. To discretize a local path (and eventually decide if it is part of the PCR, or not), we break the path into segments of equal length, such that the variation of every angular dof is less than $\pi/12$ and that of every length dof is smaller than 0.5Å. The path weights are assigned as described in subsection 2.2.

4.3. Experimental Results

After generating PCRs, we performed the following queries:

1) Compute the minimum-weight path between an arbitrary start conformation and the folded conformation.

2) Compute M near-optimal paths between the same start conformation and the folded conformation.

For 2), we computed 199 near-optimal paths between the start and goal conformations using the algorithm in [NB94]. For each milestone m in the best path, we computed the number of near-optimal paths that contain m or a milestone close to m.

We considered two proteins previously analyzed in [STD95]: 1hdd and 1le2. We obtained the description of the secondary structure from DSSP [KS83]. We took extra samples around the folded structure. The PCRs were constructed with 1500 to 5,000 milestones. On a 400-MHz Pentium II processor, the average PCR construction time ranged from 7 minutes for smaller roadmaps to 10 hours for larger ones. In all runs, the largest connected component contained more than 95% of all the nodes.

In the initial random sampling, each prismatic dof was uniformly assigned values between 0.5 and 6Å per amino acid on the loop. Each revolute dof was uniformly distributed in $[0,\pi]$ and each dihedral and twist dof was uniformly distributed in $[-\pi,\pi]$. To generate extra samples around a given milestone, each angle was picked within $\pm\pi/6$ of the corresponding angle in the milestone and each length was picked within ±0.5Å of the corresponding length. Figure 7 show the results for 1hdd. Energy vs. RMSD distribution is shown in (a); energy profile, rmsd profile, and the ratio of the number of times the milestones on the best path are also visited in the near-optimal paths.are shown in (b) and (c) For (a), red points correspond to samples taken around the native structure, whereas blue points are regularly sampled milestones.

For 1hdd, Figure 7 shows that there are milestones of lower energy than the folded state. This may be due to the various approximations made in the energy model. In (b), the energy profile shows a barrier just before reaching the folded state, similar to the profiles in [SA00]. But several runs led different plots, and (c) shows another profile, for a different PCR of the same size and for another random starting configuration. No barrier is observed before reaching the folded state and no node is visited extensively in the 200 best paths.

For 1le2, our PCR found a configuration which is visited in all 200 best paths. This configuration is displayed in Figure 8,
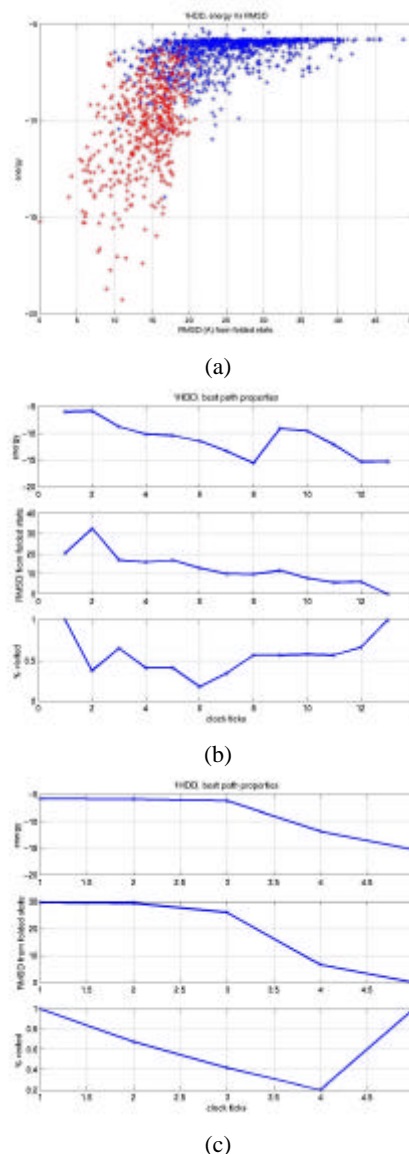

(a)

(b)

(c)

Figure 7: Results for 1hdd

along with the folded structure. Only the backbone atoms are shown. The solid vectors stand for α–helices, whereas the dashed vectors stand for loops. Both structures have the same topology, but the folded state is slightly more compact.

5. Current Research

We are pursuing our work on applying PCRs to ligand-protein binding and to protein folding. The kinematic and energetic models of the molecules need to be improved, especially for protein folding. Sampling techniques incorporating domain-specific heuristics must also be developed to take advantage of the most recent knowledge about biomolecular interactions. Concurrently, we are investigating new general techniques to produce pertinent roadmaps more quickly. Indeed, it is clear from our current work that large PCRs made of several 100,000 milestones, or more, will eventually have to be computed. Moreover, we believe that capturing function landscapes over high-dimensional spaces is a problem arising in several applications and that probabilistic roadmaps similar to PCR are a promising tool to do it. Hence, any general improvement in
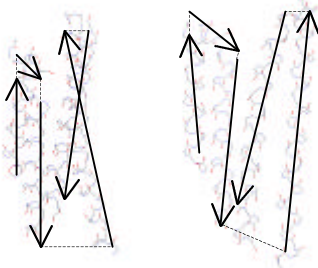
Figure 8. Folded state (l) and most visited state (r) in 200 best paths for 1LE2

computing PCRs can have a significant impact both in computational biology and beyond.

A remaining challenge in applying probabilistic roadmaps to robot motion planning is the so-called "narrow passage" issue [HKL98]. This issue also arises in protein folding, but with much greater acuity. The low-energy subset of the protein's conformation space tends to form a maze of very narrow passages [LGH97], with the (possibly unknown) folded conformation lying in one of them and the initial, extended conformation lying outside the maze. The technique proposed in [HKL98] for robot motion planning is to widen the narrow passages by allowing a small penetration of the robot into the obstacles In protein folding, using an energy function E' approximating the function E with a smaller domain of high values (see Subsection 2.4) has a similar effect. Investigating narrow passages for protein folding may eventually benefit robot motion planning.

References

[AM93] S. Arya and D.M. Mount. Approximate Nearest Neighbor Searching. *Proc. 4th Ann. ACM-SIAM Symp. on Discrete Algorithms,* 271-280, 1993.

[BSA00] O.B. Bayazit, G. Song and N.M. Amato. *Ligand Binding with OBPRM and Haptic User Input: Enhancing Automatic Motion Planning with Virtual Touch*. TR00-025, Dept. of Comp. Sci., Texas A&M U., Oct. 2000.

[DK98] Y. Duan and P.A. Kollman. Pathways to a Protein-Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous-Solution. *Science*, 282:740-744, 1998.

[FHK96] P.W. Finn, D. Halperin, L.E. Kavraki, J.C. Latombe, R. Motwani, C. Shelton, and S. Venkatasubramanian. Geometric Manipulation of Flexible Ligands. In *Lecture Notes in Comp. Sc.,* 1148, M.C. Lin and D. Manocha (eds.), Springer, NY, 67-78, 1996.

[Hai92] J.M. Haile. *Molecular Dynamics Simulation: Elementary Methods.* Wiley, NY, 1992.

[HKL98] D. Hsu, L. Kavraki, J.C. Latombe, R. Motwani, and S. Sorkin. On Finding Narrow Passages with probabilistic Roadmap Planners. In *Robotics: The Algorithmic Perspective*, P.K. Agarwal, L.E. Kavraki, and M.T. Mason (eds.), A K Peters, Natick, MA,. 141-153, 1998.

[HLMK99] D. Hsu, J.C. Latombe, R. Motwani, and L.E. Kavraki. Capturing the Connectivity of High-Dimensional Geometric Spaces by Parallelizable Random Sampling Techniques. In *Advances in Randomized Parallel Computing*, P.M. Pardalos and S. Rajasekaran (eds.), Combinatorial Optimization Series, Kluwer, Boston, MA, 159-182, 1999.

[KS83] W. Kabsch and C. Sander. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymer*, 22(12):2577-637, 1983.

[KSLO96] L.E. Kavraki, P. Svetska, J.C. Latombe, and M. Overmars. Probabilistic Roadmaps for Path Planning in High-Dimensional Configuration Spaces. *IEEE Tr. Rob. and Autom.*, 12(4):566-580, 1996.

[KB99] S.W. Kim and D. Boley. *Building and Navigating a Network of Local Minima.*. TR 99-033, CSE Dept., U. of Minnesota, 1999.

[KBK94] R.M. Knegtel, R. Boelens, and R. Kaptein. Monte Carlo Docking of Protein-DNA Complexes: Incorporation of DNA Flexibility and Experimental Data. *Protein Eng.* 7(6), 761-7, 1994.

[KL99] P. Koehl and M. Levitt. De Novo Protein Design. I. In Search of Stability and Specificity. *J. Mol. Biol.*, 293:1161-1181, 1999.

[LFKL00] S.M. LaValle, P.W. Finn, L.E. Kavraki, and J.C. Latombe. A Randomized Kinematics-Based Approach to Pharmacophore-Constrained Conformational Search and Database Screening. *J. of Comp. Chem.*, 21(9):731-747, 2000.

[LGH97] M. Levitt, M. Gerstein, E. Huang, S. Subbiah, and J. Tsai. Protein Folding: The Endgame. *Ann. Rev. Biochem.*, 66:549-579, 1997.

[MM97] C. Mata and J. Mitchell. A New Algorithm for Computing Shortest Paths in Weighted Planar Subdivisions. *Proc. 13th Int. Annual Symp. On Comp. Geom.*, ACM Press, New York, 264-273, 1997.

[NB94] D. Naor and D.L. Brutlag.. On Near-Optimal Alignments of Biological Sequences. *J. Comp. Biol.,* 1(4):349-366, 1994.

[PR97] V.S. Pande and D.S. Rokhsar. Is the Molten Globule a Third Phase of Proteins? *Proc. of the Nat. Acad. of Science*, USA, 1997.

[SH90] K. Sharp and B. Honig, Electrostatic interactions in macromolecules: theory and applications. *Ann. Rev. Biophys. Chem.,* 19:301-32, 1990

[SLB99] A.P. Singh, J.C. Latombe, and D.L. Brutlag. A Motion Planning Approach to Flexible Ligand Binding. *Proc. 7th Int. Conf. on Intel. Sys. for Mol. Bio.*, AAAI Press, Menlo Park, CA, 252-261, 1999.

[SK93] B.K. Shoichet and I.D. Kuntz. Matching Chemistry and Shape in Molecular Docking. *Protein Eng.* 6(7) 723-32, 1993.

[SB97] A.P. Singh and D.L. Brutlag. Hierarchical Protein Structure Superposition Using Both Secondary Structure and Atomic Representations. *Proc. 5th Int. Conf. On Intell. Syst. for Mol. Bio.),* AAAI Press, Menlo Park, CA, 284-293, 1997.

[SA00] G. Song and N.M. Amato. *Using Motion Planning to Study Protein Folding Pathways*. TR00-026, Dept. of Comp. Sci., Texas A&M U., Oct. 2000.

[STD95] S. Sun, P.D. Thomas, and K.A. Dill. A Simple Protein Folding Algorithm Using a Binary Code and Secondary Structure Constraints. *Protein Eng.* 8:769-778, 1995.

[TPK00] M. Teodoro, G. Phillips, and L.E. Kavraki. Singular Value Decomposition of Protein Conformational Motions: Application to HIV-1 Protease. *Currents in Comp. Mol. Bio.*, M. Satoru, R. Shamir, and T. Tagaki (eds.), Universal Academy Press., 198-199, 2000.