# Simulating Protein Motions with Rigidity Analysis[⋆]

Shawna Thomas, Xinyu Tang, Lydia Tapia, and Nancy M. Amato

Parasol Lab, Dept. of Comp. Sci., Texas A&M University, College Station, TX 77843

**Abstract.** Protein motions, ranging from molecular flexibility to large-scale conformational change, play an essential role in many biochemical processes. Despite the explosion in our knowledge of structural and functional data, our understanding of protein movement is still very limited. In previous work, we developed and validated a motion planning based method for mapping protein folding pathways from unstructured conformations to the native state. In this paper, we propose a novel method based on rigidity theory to sample conformation space more effectively, and we describe extensions of our framework to automate the process and to map transitions between specified conformations. Our results show that these additions both improve the accuracy of our maps and enable us to study a broader range of motions for larger proteins. For example, we show that rigidity-based sampling results in maps that capture subtle folding differences between protein G and its mutations, NuG1 and NuG2, and we illustrate how our technique can be used to study large-scale conformational changes in calmodulin, a 148 residue signaling protein known to undergo conformational changes when binding to $Ca^{2+}$. Finally, we announce our web-based protein folding server which includes a publically available archive of protein motions: http://parasol.tamu.edu/foldingserver/

## 1 Introduction

Protein motions, ranging from molecular flexibility to large-scale conformational change, play an essential role in many biochemical processes. For example, conformational change often occurs in binding. While no consensus has been reached regarding models for protein binding, the importance of protein flexibility in the process is well established by the ample evidence that the same protein can exist in multiple conformations and can bind to structurally different molecules.

Our understanding of molecular movement is still very limited and has not kept pace with the explosion of knowledge regarding protein structure and function. There are several reasons for this. First, the structural data in repositories like the Protein Data Bank (PDB) [8] consists of the spatial coordinates of each

[⋆] Supported in part by NSF Grants EIA-0103742, ACR-0081510, ACR-0113971, CCR-0113974, ACI-0326350, and by the DOE. Thomas supported in part by an NSF Graduate Research Fellowship.

atom. Unfortunately, the experimental methods used to collect this data cannot operate at the time scales necessary to record detailed large-scale protein motions. Second, traditional simulation methods such as molecular dynamics and Monte Carlo methods are computationally too expensive to simulate long enough time periods for anything other than small peptide fragments.

There has been some attention focused on methods for modeling protein flexibility and motion. One notable effort is the Database of Macromolecular Movements [15, 14]. They generate and archive protein 'morphs' that interpolate between two different protein conformations. While the method used is more chemically realistic than straight-line interpolation (as described in Section 2), it was selected over other more accurate methods for computational efficiency and is known to have problems for some kinds of large deformations.

In previous work [3, 2, 42, 41], we developed a new computational technique for studying protein folding that builds an approximate map of a protein's potential energy landscape. This map contains thousands of feasible folding pathways to the known native state enabling the study of global landscape properties. We obtained promising results for several small proteins (60–100 amino acids) and validated our pathways by comparing secondary structure formation order with known experimental results [3].

**Our Contribution.** We augment our framework with three powerful new concepts that enable us to study a broader range of motions for larger proteins:

- We propose a new method based on rigidity theory to sample conformations.
- We generalize our PRM framework to map specified transitions.
- We present a new framework to automate the map building process.

Our new rigidity-based sampling allows us to study larger proteins by more efficiently characterizing the protein's energy landscape with fewer, more realistic conformations. We exploit rigidity information by focusing sampling on (currently) flexible regions. This results in smaller, better maps. In one dramatic case study, we show that rigidity-based sampling and analysis reveals the folding differences between protein G and its mutants, NuG1 and NuG2, which is an important 'benchmark' set that has been developed by the Baker Lab [36].

Extending our framework to focus on particular conformations enables us to investigate questions related to the transition between particular conformations, e.g., when studying folding intermediates, allostery, or misfolding. We provide evidence that the transitions mapped by our approach are more realistic than those provided by the computationally less expensive Morph Server [14], especially for transitions requiring large conformational changes.

The accuracy of our approach heavily depends on how densely we sample the conformation space. Previously, this was user specified and fixed. Here, we use an extension of our basic technique which incrementally samples the conformation space at increasingly denser resolution until our map of the landscape stabilizes.

Finally, we announce our protein folding server which uses our technique to generate protein transitions to the native state or between selected conformations. We invite the community to help enrich our publicly available database by submitting to our server: http://parasol.tamu.edu/foldingserver/

| Approach | Landscape | # Paths | Path Quality | Computation | Native Required |
|---|---|---|---|---|---|
| Molecular Dynamics | No | 1 | Good | Long | No |
| Monte Carlo | No | 1 | Good | Long | No |
| Statistical Model | Yes | 0 | N/A | Fast | Yes |
| **PRM-Based** (Our Approach) | Yes | Many | Approx | Fast | Yes |
| Lattice Model | Not used on real proteins | | | | |

**Table 1.** Comparison of protein motion models.

## 2 Related Work

**Protein Motion Models** Several computational approaches have been used to study protein motions and folding, see Table 1. These include lattice models [10], energy minimization [30, 44], molecular dynamics [29, 16], and Monte Carlo methods [13, 26]. Molecular dynamics and Monte Carlo methods provide a single, high quality transition pathway, but each run is computationally intensive. Statistical mechanical models [35, 1, 6], while computationally efficient, are limited to studying global averages of the energy landscape and kinetics and are unable to produce individual pathways.

**Computing Macromolecular Motions.** Gerstein et al. have developed the Database of Macromolecular Movements [15, 14] to classify protein motions. Their server produces a 'morph' movie between two target conformations in just a few minutes on a desktop PC. Their database currently includes more than 240 distinct motions.

To 'morph' between two target conformations, they first perform an alignment. Then, an iterative 'sieve-fit' procedure produces a superposition of the target conformations. The superimposed conformations are 'morphed' by interpolating the C$\alpha$ atom positions. Each intermediate conformation is energy minimized. This interpolation method, called adiabatic mapping, was selected because it has modest computational requirements yet produces chemically reasonable 'morphs.' Adiabatic mapping, however, is not guaranteed to produce accurate trajectories and in fact cannot model many large deformations.

**Motion Planning and Molecular Motions.** The motion planning problem is to find a valid path for a movable object from a start to a goal. The probabilistic roadmap method (PRM) [23] has been highly successful in solving high degree of freedom (dof) problems.

PRMs first sample random points in the movable object's conformation space (C-space). C-space is the set of all possible positions and orientations of the movable object, valid or not. Only those samples that meet feasibility requirements (e.g., collision free or low potential energy) are retained. Neighboring samples are connected to form a graph (or roadmap) using some simple local planner (e.g., a straight line). This roadmap can then be used to find the motion between different start and goal pairs by connecting them to the roadmap and extracting a path, if one exists. PRMs are simple to apply, even for high dof problems, only requiring the ability to generate random samples in C-space and test them feasibility.

PRMs have been applied to model molecular motions by modeling the molecule as an articulated linkage and replacing the typical collision detection validity

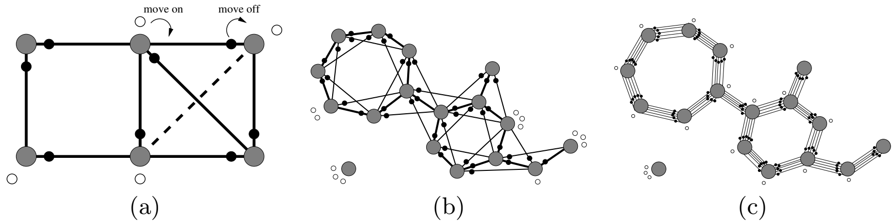


**Fig. 1.** (a) The result of the pebble game on a 2D graph. Pebbles may be free (white) or covering (black). Constraints are marked as independent (solid) or redundant (dashed). Pebbles may be rearranged as shown. Rigidity models for a sample molecule: (b) bar-joint and (c) body-bar.

check with some measure of physical viability (e.g., potential energy). Singh, Latombe and Brutlag first applied PRMs to protein/ligand binding [40]. In subsequent work, our group used another PRM variant on this problem [7]. Our group was the first to adapt PRMs to model protein folding pathways [3, 2, 42, 41]. Apaydin et. al. [5, 4] also applied PRMs to proteins, however their work differs from ours in several aspects. First, they model the protein at a much coarser level, considering all secondary structure elements in the native state to already be formed and rigid. Second, while our focus is on studying the transition process, their focus has been to compare the PRM approach with other computational methods such as Monte Carlo simulation. Cortes and Simeon used a PRM-based approach to model long loops in proteins [12]. Recently, we adapted the PRM framework to study RNA folding kinetics [45].

**Rigidity Theory and Protein Flexibility.** Several computational approaches study protein rigidity and flexibility. One approach infers rigidity and flexibility by comparing different known conformations [37, 9]. Molecular dynamics has been used to extract flexibility information from simulated motion [32, 11, 24]. A third method studies rigidity/flexibility of a single conformation [21, 22, 33]. Here, we use a rigidity analysis technique belonging to the third class of approaches called the pebble game [19, 18] to better simulate motion. It is fast and efficient; we can apply it to every conformation we sample.

The pebble game is a constraint counting algorithm which determines the dof in a two-dimensional graph, along with its rigid/flexible regions. In 2D, the pebble game assigns each vertex two pebbles, representing its two dof, see Figure 1a. Each edge/constraint is examined to determine if it is independent or redundant. If two free pebbles can be placed on both endpoints of the edge, then it is marked independent and covered by a pebble from one of its incident vertices. Once an edge is covered by a pebble, it remains covered, although which vertex the pebble comes from may change. Pebbles may be rearranged as shown in Figure 1a. If pebbles cannot be rearranged to get two free pebbles on both of an edge's endpoints, then the edge is marked redundant and indicates a rigid region. In the end, the remaining free pebbles indicate the graph's dof.

The 2D pebble does not generalize to 3D for arbitrary graphs, but it can be applied to 3D bond-bending networks [18]. A bond-bending network is a truss structure with constraints between nearest neighbors and next-nearest neighbors. A protein, with fixed bond lengths and bond angles, can be modeled as a bond-bending network where atoms are modeled as vertices with 3 dof and bonds are modeled as edges, called the bar-joint model, see Figure 1b. It has been successfully used by several applications to study protein rigidity and flexibility [20, 39, 17, 28]. An alternative model, the body-bar model, represents atoms as rigid bodies with 6 dof and the torsional bonds between them as 5 bars/constraints [46], see Figure 1c. Both models are conjectured to be equivalent [18].

## 3 Modeling Molecular Motions with PRMs

We have successfully applied the PRM framework to study protein folding pathways [3, 2, 42, 41]. We model the protein as an articulated linkage. Using a standard modeling assumption for proteins that bond angles and bond lengths are fixed [43], the only dof in our model are the backbone's phi and psi torsional angles which are modeled as revolute joints with values $[0, 2\pi)$.

The strategy follows the general PRM methodology sketched in Section 2. First, different protein conformations are sampled. A sample $q$, with potential energy $E(q)$, is accepted with the probability:

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

where $E_{\min}$ is the potential energy of the open chain and $E_{\max}$ is $2E_{\min}$. Next, node connection is done in the same way as traditional PRMs except that each connection is assigned a weight to reflect its energetic feasibility. The weight for the edge $(q_1, q_2)$ is a function of all the intermediate conformations along the edge $\{q_1 = c_0, c_1, \ldots, c_{n-1}, c_n = q_2\}$. For each pair of consecutive conformations $c_i$ and $c_{i+1}$, the probability $P_i$ of transitioning from $c_i$ to $c_{i+1}$ depends on the difference in their potential energies $\Delta E_i = E(c_{i+1}) - E(c_i)$:

$$P_i = \begin{cases} e^{\frac{-\Delta E_i}{kT}} & \text{if } \Delta E_i > 0 \\ 1 & \text{if } \Delta E_i \leq 0 \end{cases}$$

This keeps the detailed balance between two adjacent states, and enables the weight of an edge to be computed by summing the logarithms of the probabilities for consecutive pairs of conformations in the sequence. Edge weights are not transition rates, but the logarithm of transition rates. This enables edge weights to follow the summation rule (instead of the multiplication rule for transition rates) and facilitates the use of graph algorithms to extract shortest paths.

The samples and connections form a roadmap from which we can typically extract thousands of transition pathways. With our original method, in just a few hours on a desktop PC, we obtained promising results for many small proteins (60–100 residues) and validated our pathways by comparing the secondary structure formation order with known experimental results [3]. In one case study, our technique was sensitive enough to identify folding differences for structurally similar proteins G and L [42].

**Potential Energy Calculations.** As in our previous work, we use a coarse potential function similar to [29]. We use a step function approximation of the van der Waals component and model all side chains as equal radii spheres with zero dof. If two spheres are too close (i.e., $< 2.4$Å during sampling and 1.0Å during connection), a very high potential is returned. Otherwise the potential is:

$$U_{tot} = \sum_{restraints} K_d\{[(d_i - d_0)^2 + d_c^2]^{1/2} - d_c\} + E_{hp}$$

where $K_d$ is 100 kcal/mol and $d_0 = d_c = 2$ Å as in [29]. The first term represents constraints favoring known secondary structure through main-chain hydrogen bonds and disulphide bonds, and the second term is the hydrophobic effect. The hydrophobic effect is computed as follows: if two hydrophobic residues are within 6Å of each other, then the potential is decreased by 100 kJ/mol.

### 3.1 Rigidity-Based Sampling

The roadmap produced by our technique is an approximation of the protein's energy landscape. Roadmap quality is measured both by how realistic (as compared to experimental data) its pathways are and by how many samples are required to achieve the desired accuracy. The latter is important because it determines what size molecules can be analyzed.

Hence, sampling is the key to producing a good approximation of the landscape. Note that only a relatively small portion of the conformation space 'near' the target conformation(s) is of interest in modeling motions. This implies that we should use biased sampling to cover the regions of interest efficiently.

In previous work [3, 2, 42, 41], we obtained a denser distribution of samples near the target conformation through an iterative sampling process where we apply small Gaussian perturbations to existing conformations, beginning with the target conformation. This approach still requires many samples (e.g., 10,000) for relatively small proteins (e.g., 60–100 residues). To apply our method to larger proteins, we need strategies to generate 'better' samples; they should be more physically realistic and represent 'stepping stones' for conformational transitions.

In this work, we follow the same strategy as before, but use rigidity analysis to restrict how to perturb a conformation. We first use rigidity analysis to determine which bonds are independently flexible, dependently flexible, and rigid, see Figure 2b. Independently flexible bonds can be perturbed without affecting the rest of the bonds in the system. Dependently flexible bonds form a set of bonds such that perturbing any one of these bonds results in a corresponding perturbation in the rest of the set.

If the bond is independently flexible, we perturb with a high probability, $P_{flex}$. If the bond is rigid, we perturb with a low probability, $P_{rigid}$. For each dependently flexible set, we randomly select $d$ bonds to perturb with probability $P_{flex}$ and perturb the remaining bonds with probability $P_{rigid}$, where $d$ is the internal dof in the set. Perturbing rigid dof ensures good coverage of the space.

**Rigidity Model.** We employ the body-bar model to analyze a conformation's rigidity. With the body-bar model, we can represent the protein at a residue level, a closer match to our phi-psi model for sampling than the bar-joint model with a more detailed all-atoms view.

We model the protein simply as a chain of rigid bodies, each representing one torsional dof, see Figure 2a. We model each peptide bond and disulphide bond with 5 bars, each hydrogen bond with 2 bars, and each hydrophobic contact with 1 bar. On all conformations tested, this yields the same rigid and flexible regions as the equivalent bar-joint model on an all-atoms representation of the protein.

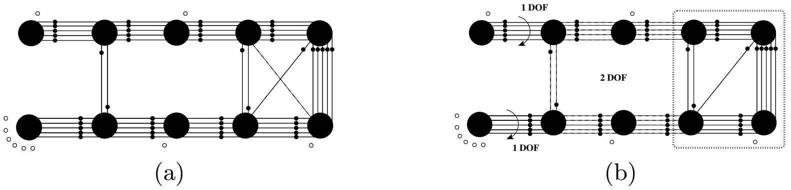


**Fig. 2.** (a) Model of a 5 residue protein. Each residue has two rigid bodies. We model peptide bonds and disulphide bonds with 5 bars, hydrogen bonds with 2 bars, and hydrophobic interactions with 1 bar. Redundant constraints (dashed lines) identified by the pebble game. (b) Pebble game results: a rigid cluster (dotted box), a 2 dof dependent hinge set (dashed lines), and independently flexible bonds (arcs).

**Rigidity Map.** We can also use rigidity analysis to define a new residue mapping and distance metric. A rigidity map, $r$, is similar to a contact map. Rigid body pairs $(i, j)$ from the rigidity model are marked if they have the same rigidity relationship: 2 if they are in the same rigid cluster, 1 if they are in the same dependent hinge set, and 0 otherwise. (Recall that there are two rigid bodies for each residue representing the two torsional dof.) Figure 3a shows the rigidity map of the native state for protein G with rigid clusters (black) and dependent hinge sets (green/shaded). Rigidity maps provide a convenient way to define a rigidity distance metric, $r_{\text{dist}}(q_1, q_2)$, between two conformations $q_1$ and $q_2$ where $n$ is the number of residues:

$$r_{\text{dist}}(q_1, q_2) = \sum_{0 \leq i < j \leq 2n} (r_{q_1}(i, j) \neq r_{q_2}(i, j)).$$

### 3.2 Automatic Roadmap Construction

Roadmap accuracy depends on the sampling density. Previously, this was user specified and difficult to tune. Here, we automate roadmap construction by building the roadmap incrementally [47]. We first build a roadmap with a low sampling density as described above. Then, we test the roadmap to see if it has stabilized as specified by a set of evaluation criteria. We continue to augment the roadmap with more samples and connections until it satisfies the evaluation criteria. This provides two key advantages over our previous work: (1) the roadmap is constructed automatically at the appropriate resolution, and (2) we reuse all previous computation reducing runtime cost by several factors.

For protein folding, we build a roadmap until the secondary structure formation order along its pathways stabilizes. A piece of secondary structure is 'formed' when the distance between its rigidity map (defined in Section 3.1) and that piece's rigidity map in the target state is within 0.8, normalized to the range [0,1]. The pathway's secondary structure formation order is then the order at which pieces are 'formed.' We examine every pathway in the roadmap from an unstructured conformation to the target state and group them by this ordering. We consider the roadmap stable when the percentage of each group does not vary from the previous roadmap by more than 10%.

### 3.3 Mapping Specified Transitions

We extend our PRM framework to study specific large-scale conformational changes by iteratively sampling around each target conformation and connecting samples together as described earlier in Section 3. Thus our roadmaps contain the target conformations, as well as the transitions between them, and approximate the energy landscape encompassing the transition under study.

We can study problems such as transitions between known folding intermediates, transitions between bound and unbound conformations to a ligand, misfolded proteins, and allostery interactions. For example, several devastating diseases such as scrapie in sheep and goats, bovine spongiform encephalopathy (Mad Cow disease), and Creutzfeldt-Jakob disease in humans are caused by misfolded proteins called prions [38]. Insight into how these proteins misfold could help develop better drugs.

To map specific large-scale transitions, we interleave sampling and connection to incrementally build a roadmap as in Section 3.2. The only difference here is we sample around each target conformation (as in Section 3.1) during each round of roadmap construction. Then we connect samples together and compute edge weights as before. We continue until the roadmap adequately represents the protein's energy landscape near the target conformations and between them. From this roadmap, we can extract multiple low energy transition pathways between target conformations and characterize the energy barriers between them.

We build the roadmap until the maximum network flow between each target conformation pair is above a threshold. For maximum network flow, edges are assigned a capacity, and the goal is to determine how much flow can be achieved between two points in the graph. Here, we define edge capacity as the inverse of the edge weight. Thus, the maximum network flow between two conformations approximates the transition rate between them [27].

## 4 Results and Discussion

We investigate the ability of our rigidity-based sampling strategy to efficiently sample the protein's conformation space. We also look at examples of large-scale conformational change between specific target states for several small proteins and compare our results with 'morphs' from the Database of Macromolecular Movements [15, 14]. In all experiments, we set $P_{flex}$ to 0.8 and $P_{rigid}$ to 0.2. We use a straight line local planner and attempt to connect each conformation with its 50 nearest neighbors. We measure distance between two conformations as the difference between their rigidity maps (see Section 3.1).

**Improved Sampling.** Rigidity analysis coupled with automatic roadmap construction greatly improves the efficiency of our PRM framework by restricting the sample space in a physically realistic way. We can build smaller roadmaps that better reflect the landscape. We built roadmaps for several previously studied proteins [2, 41]. For each protein, we compare our new automatic framework with rigidity-based sampling to our previous sampling technique with fixed sampling density. Table 2 shows the roadmap size and connectivity from both methods. Both methods give the same secondary structure formation order distribution. When available, these results also indicate the same dominant secondary structure formation order seen in experiment [31]. In all cases, the rigidity-based roadmaps produce equivalent folding pathways as the previous method with smaller, more efficient roadmaps and increases connectivity. Thus, we can study much larger proteins than before.

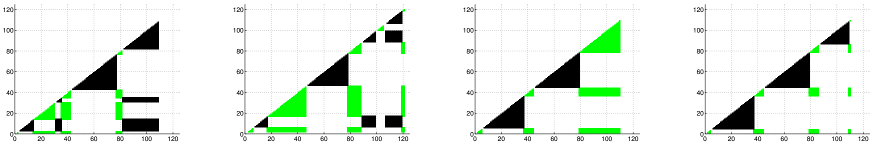
| PDB | | | Gaussian Sampling | | | | Rigidity Sampling | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Identifier | Length | Structure | Nodes | Edges | $N+E$ | $E/N$ | Nodes | Edges | $N+E$ | $E/N$ |
| 1AB1 | 46 | $2\alpha+2\beta$ | 24206 | 386974 | 411180 | 15.99 | 6000 | 158286 | 164286 | 26.38 |
| 1CCM | 46 | $1\alpha+3\beta$ | 43646 | 728964 | 772610 | 16.70 | 10000 | 456080 | 466080 | 45.61 |
| 1RDV | 52 | $2\alpha+3\beta$ | 33691 | 457392 | 491083 | 13.58 | 4000 | 166702 | 170702 | 41.68 |
| 1EGF | 53 | $3\beta$ | 27356 | 391146 | 418502 | 14.30 | 4000 | 164902 | 168902 | 41.23 |
| 1PRB | 53 | $5\alpha$ | 44551 | 696708 | 741259 | 15.64 | 4000 | 126562 | 130562 | 31.64 |
| 1SMU | 54 | $3\alpha+3\beta$ | 35501 | 557416 | 592917 | 15.70 | 4000 | 158852 | 162852 | 39.71 |
| 1FCA | 55 | $2\alpha+4\beta$ | 38216 | 489840 | 528056 | 12.82 | 4000 | 162526 | 166526 | 40.63 |
| 1VGH | 55 | $1\alpha+4\beta$ | 38216 | 631936 | 670152 | 16.54 | 4000 | 157454 | 161454 | 39.36 |
| 1GB1 | 56 | $1\alpha+4\beta$ | 34236 | 912908 | 947144 | 26.66 | 4000 | 160552 | 164552 | 40.14 |
| 1SHG | 57 | $5\beta$ | 24696 | 270232 | 294928 | 10.94 | 18000 | 654884 | 672884 | 36.38 |
| 1BPI | 58 | $2\alpha+2\beta$ | 28426 | 399418 | 427844 | 14.05 | 4000 | 112010 | 116010 | 28.00 |
| 4PTI | 58 | $2\alpha+2\beta$ | 39121 | 389468 | 428589 | 9.96 | 4000 | 160100 | 164100 | 40.03 |
| 1HCC | 59 | $7\beta$ | 33691 | 453628 | 487319 | 13.46 | 28000 | 1079904 | 1107904 | 38.57 |
| 1BDD | 60 | $3\alpha$ | 58486 | 888298 | 946784 | 15.19 | 6000 | 195950 | 201950 | 32.66 |
| 1TCP | 60 | $2\alpha+2\beta$ | 32786 | 354262 | 387048 | 10.81 | 4000 | 163692 | 167692 | 40.92 |
| 2ADR | 60 | $2\alpha+2\beta$ | 42723 | 701942 | 744665 | 16.43 | 8000 | 339498 | 347498 | 42.44 |
| 2PTL | 62 | $1\alpha+4\beta$ | 23921 | 281334 | 305255 | 11.76 | 4000 | 159728 | 163728 | 39.93 |
| 1COA | 64 | $1\alpha+5\beta$ | 27746 | 403438 | 431184 | 14.54 | 4000 | 160838 | 164838 | 40.21 |
| 2CI2 | 65 | $2\alpha+5\beta$ | 27746 | 389670 | 417416 | 14.04 | 8000 | 228706 | 236706 | 28.59 |
| 1NYF | 67 | $5\beta$ | 23921 | 262376 | 286297 | 10.97 | 6000 | 249450 | 255450 | 41.58 |
| 1MJC | 69 | $7\beta$ | 23481 | 226942 | 250423 | 9.66 | 4000 | 153140 | 157140 | 38.29 |
| 1HOE | 74 | $7\beta$ | 30626 | 184012 | 214638 | 6.01 | 4000 | 103668 | 107668 | 25.92 |
| 1UBQ | 76 | $1\alpha+5\beta$ | 25206 | 236216 | 261422 | 9.37 | 4000 | 154192 | 158192 | 38.55 |
| 1O6X | 81 | $2\alpha+3\beta$ | 40931 | 342138 | 383069 | 8.36 | 4000 | 133544 | 137544 | 33.39 |
| 1PBA | 81 | $4\alpha+3\beta$ | 26476 | 203974 | 230450 | 7.70 | 8000 | 282960 | 290960 | 35.37 |
| 2ABD | 86 | $5\alpha$ | 27956 | 681796 | 709752 | 24.39 | 18000 | 953900 | 971900 | 52.99 |

**Table 2.** Comparison of rigidity-based sampling to previous work for several proteins. In all cases, rigidity-based sampling significantly reduces the required roadmap size ($N+E$) to produce equivalent pathways. It also increased roadmap connectivity ($E/N$).

**Case study of proteins G, L NuG1, and NuG2.** Proteins G, L, and mutants of protein G, NuG1 and NuG2 [36], present a good test case for our technique because they are known to fold differently despite having similar structure. All proteins are composed of a central $\alpha$-helix and a 4-stranded $\beta$-sheet: $\beta$

strands 1 and 2 form the N-terminal hairpin ($\beta$1-2) and $\beta$ strands 3 and 4 form the C-terminal hairpin ($\beta$3-4). Native state out-exchange experiments and pulse labeling/competition experiments for proteins G and L indicate that $\beta$1-2 forms first in protein L, and $\beta$3-4 forms first in protein G [31]. This is consistent with $\Phi$-value analysis on G [34] and L [25]. In [36], protein G is mutated in both hairpins to increase the stability of $\beta$1-2 and decrease the stability of $\beta$3-4. $\Phi$-value analysis indicates that the hairpin formation order for both NuG1 and NuG2 is switched from the wild type.

Our previous sampling strategy [42] was able to capture the folding differences between proteins G and L, but not between protein G and NuG1 or NuG2. Our new rigidity-based sampling and analysis is able to also capture the correct folding behavior of NuG1 and NuG2, see Table 3. In addition, our rigidity-based technique can also help to explain the stability shift in NuG1 and NuG2. For example, consider their native state rigidity maps shown in Figure 3. In all four proteins, the central alpha helix remains completely rigid. We also see increased rigidity in $\beta$1-2 from protein G to NuG1 and NuG2 as suggested in [36].



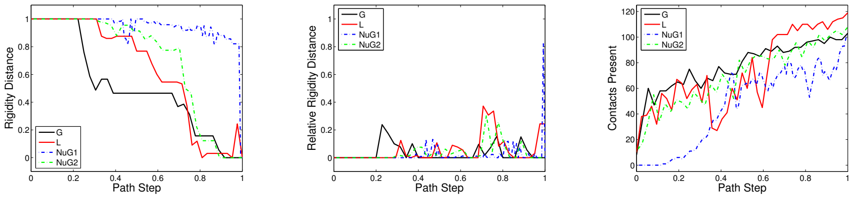
(a) (b) (c) (d)

**Fig. 3.** Rigidity maps for the native states of proteins (a) G, (b) L, (c) NuG1, and (d) NuG2. Rigid clusters are black and dependent hinge sets are shaded/green.

| Protein | Experimental Formation Order | Rigidity Formation Order | % |
|---|---|---|---|
| G | [$\alpha$,$\beta$1,$\beta$3,$\beta$4], $\beta2^1$ [$\alpha$,$\beta$4], [$\beta$1,$\beta$2,$\beta$3]$^2$ | $\alpha$, $\beta$3-4, $\beta$1-2 | 99.4 |
| L | [$\alpha$,$\beta$1,$\beta$2,$\beta$4], $\beta3^1$ [$\alpha$,$\beta$1], [$\beta$2,$\beta$3,$\beta$4]$^2$ | $\beta$1-2, $\alpha$, $\beta$3-4 | 100.0 |
| NuG1 | $\beta$1-2, $\beta$3-4$^3$ | $\alpha$, $\beta$1-2, $\beta$3-4 | 97.6 |
| | | $\beta$1-2, $\alpha$, $\beta$3-4 | 1.6 |
| NuG2 | $\beta$1-2, $\beta$3-4$^3$ | $\alpha$, $\beta$1-2, $\beta$3-4 | 96.6 |
| | | $\beta$1-2, $\alpha$, $\beta$3-4 | 1.1 |
| | | $\beta$3-4, $\beta$1-2, $\alpha$ | 1.1 |

**Table 3.** Comparison of secondary structure formation orders for proteins G, L, NuG1, and NuG2 with known experimental results: $^1$hydrogen out-exchange experiments [31], $^2$pulsed labeling/competition experiments [31], and $^3\Phi$-value analysis [36]. Brackets indicate no clear order. In all cases, our technique predicted the secondary structure formation order seen in experiment. Only formation orders greater than 1% are shown.

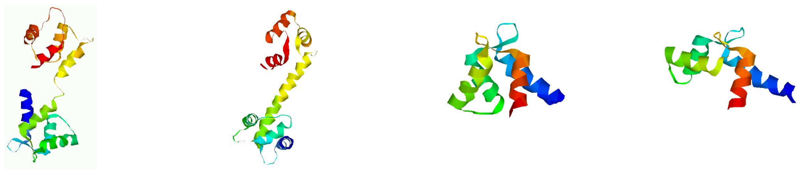
We can also use rigidity-based analysis to study dynamic changes along a folding pathway, see Figure 4. We see a distinction between the profiles for protein G where $\beta$3-4 forms first and the others where $\beta$1-2 forms first. For protein G, the rigidity profile (a) shows a plateau halfway along the folding pathway, where the others do not. Protein G (b) also exhibits larger changes in rigidity earlier in the pathway while the others exhibit larger changes later.



(a) (b) (c)

**Fig. 4.** Folding pathway profiles for proteins G, L, NuG1, and NuG2: (a) rigidity distance to the target state, (b) relative rigidity distance to the target state, and (c) contacts present. There is a distinction between the profiles for protein G where $\beta$3-4 forms first and the others where $\beta$1-2 forms first.

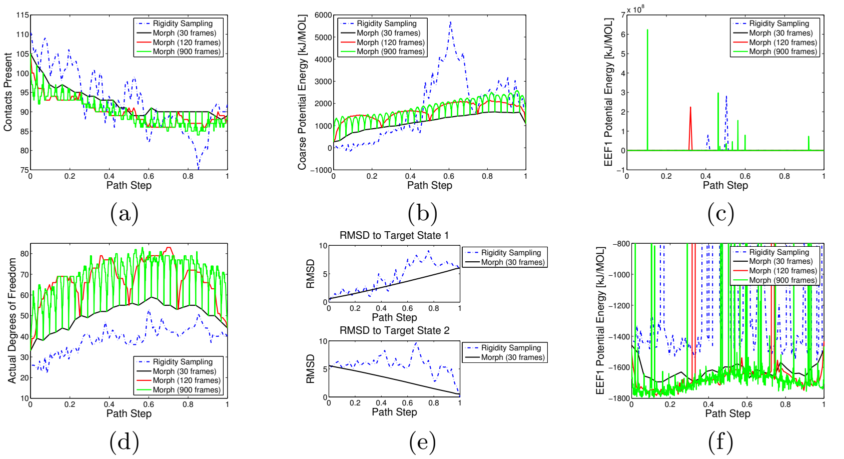
**Large-Scale Conformational Change.** Calmodulin is a 148-residue signaling protein that binds to Ca$^{2+}$ to regulate several processes in the cell. It is composed of 4 EF-hands joined by a flexible central $\alpha$ helix. When binding to Ca$^{2+}$, it undergoes two large-scale conformational changes: (1) the central $\alpha$ helix unravels to bring the protein from a dumbbell conformation to a more globular conformation (Figure 5a–b) and (2) the $\alpha$ helices in each domain reorganize (Figure 5c–d).



(a) (b) (c) (d)

**Fig. 5.** Conformational changes of calmodulin: (a) calcium-free state (1CFD) to (b) bound state (1CLL) and of the N-terminal domain: (c) calcium-free to (d) bound.

We built a roadmap biased towards both target states as outlined in Section 3.3. Figure 6 compares pathway profiles of the most energetically feasible transition between the two states in our roadmap and 'morphs' of various resolution obtained from the Morph server [15, 14]. We examined pathway profiles for energy, contacts present, dof computed by rigidity analysis, RMSD distance to the target states, and rigidity distance to the target states. Note that since the Morph server alters the original target conformations, their profile endpoints do not always align with our pathways. One striking observation is the regularity of the concavities for the 'morphs' corresponding to the various resolution levels across all the profiles except for the RMSD profiles in which the RMSD to the target states seems to change monotonically with the path step. These regularities in the 'morphs' would not be expected in actual transition pathways, e.g., one would not expect a monotonic increase in RMSD from 1CFD to 1CLL. In contrast, our roadmap pathways profiles are more plausible — they exhibit trends, but also have reasonable fluctuations. Indeed, this type of behavior has also been observed by other researchers, e.g., in [48], Monte Carlo simulations indicate a wide range of transition pathways and event durations.



(a) (b) (c)

(d) (e) (f)

**Fig. 6.** Pathway profiles for the calmodulin N-terminal domain: (a) contacts present, (b) coarse potential energy, (c,f) all-atoms potential energy, (d) dof computed by rigidity analysis, and (e) RMSD to both target states. For RMSD, only the 30 frame 'morph' shown because all resolutions are nearly identical.

Figure 6a,d shows the contacts present and dof computed by rigidity analysis along the pathway. Note that the protein does not completely unfold, but maintains a large number of contacts and loses few dof. Generally, the actual dof is inversely proportional to the number of contacts present. It is interesting to note, however, that we see a slight break in this relationship on the second half of the pathway where the peaks in dof do not match up with the peaks in number of contacts. Regions of the protein become stressed when the number of contacts increases without a corresponding decrease in dof.

We investigated several other protein transitions in a similar way, see Table 4. We measure % dof gained as the difference between the maximum dof along the pathway and the minimum dof of the starting/ending conformations, as a percentage of the total dof possible (2*length). Most transitions do not involve a complete unfolding of the protein. In fact, several have % dof gain less than 10%. We also captured different types of transitions including smooth transitions without any significant energy barriers (i.e., 1PRV, 1BMR, and 1FOX) and those with multiple energy barriers (i.e., 2VGH and 1CMF).

We also compared 'morphs' of various resolutions to our transition pathways when possible. (The Morph server was not able to produce some higher resolution 'morphs' for transitions 1BMR–1FH3 and 1PRV–1PRU.) Across all transitions, we observed the same concavity pattern phenomenon for the 'morph' transitions as seen in calmodulin (Figure 6) for energy, contacts, degrees of freedom, and rigidity distance to the targets. Here also, the RMSD to the target states essentially changed monotonically with the path step. Again, our

| Transition IDs | | Length | % Dof Gained | # Barriers |
|---|---|---|---|---|
| 2VGH | 1VGH | 55 | 21.8 | 2 |
| 1PRV | 1PRU | 56 | 5.4 | 0 |
| 1BMR | 1FH3 | 67 | 32.8 | 0 |
| 1CFD | 1CLL | 72 | 18.1 | 1 |
| 1CMF | 1CMG | 73 | 24.7 | 2 |
| 1FOX | 2FOW | 76 | 3.9 | 0 |
| 1PFH | 1HDN | 85 | 43.5 | 1 |

**Table 4.** Pathway results for transitions studied. Most do not involve large unfolding.

pathways did not exhibit these unrealistic regularities. Additional path profiles for all the transitions studied here can be found on our folding server: http://parasol.tamu.edu/foldingserver/

## 5 Conclusion

In this paper, we describe how to augment our PRM-based approach to study a broader range of motions for larger proteins. We proposed a method based on rigidity theory to sample more efficiently and to generate transitions between specified conformations. We also demonstrated that our approach yields more physically realistic transitions than those produced by the computationally less expensive Morph server. We invite the community to help enrich our publicly available motion database at http://parasol.tamu.edu/foldingserver/

## 6 Acknowledgments

We thank Ileana Streanu, Leslie Kuhn, and Walter Whitely for introducing us to rigidity theory and many insightful discussions. We also thank Luke Hunter, Kasia Leyk, and Dawen Xie for help with experiments.

## References

1. E. Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA*, 96(20):11305–11310, 1999.
2. 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. Comput. Biol.*, 10(3-4):239–256, 2003. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2002.
3. N. M. Amato and G. Song. Using motion planning to study protein folding pathways. *J. Comput. Biol.*, 9(2):149–168, 2002. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2001.
4. M. Apaydin, D. Brutlag, C. Guestrin, D. Hsu, and J.-C. Latombe. Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 12–21, 2002.
5. M. Apaydin, A. Singh, D. Brutlag, and J.-C. Latombe. Capturing molecular energy landscapes with probabilistic conformational roadmaps. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 932–939, 2001.

6. D. Baker. A surprising simplicity to protein folding. *Nature*, 405:39–42, 2000.
7. O. B. Bayazit, G. Song, and N. M. Amato. Ligand binding with OBPRM and haptic user input: Enhancing automatic motion planning with virtual touch. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 954–959, 2001. This work was also presented as a poster at *RECOMB 2001*.
8. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
9. N. Boutonnet, M. Rooman, and S. Wodak. Automatic analysis of protein conformational changes by multiple linkage clustering. *J. Mol. Biol.*, 253:633–647, 1995.
10. J. Bryngelson, J. Onuchic, N. Socci, and P. Wolynes. Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Protein Struct. Funct. Genet*, 21:167–195, 1995.
11. D. Case. Molecular dynamics and normal mode analysis of biomolecular rigidity. In M. Thorpe and P. Duxbury, editors, *Rigidity theory and applications*, pages 329–344. Kluwer Academic/Plenum Publishers, 1999.
12. J. Cortes, T. Simeon, M. Remaud-Simeon, and V. Tran. Geometric algorithms for the conformational analysis of long protein loops. *J. Computat. Chem.*, 25, 2004.
13. D. Covell. Folding protein $\alpha$-carbon chains into compact forms by Monte Carlo methods. *Proteins: Struct. Funct. Genet.*, 14(4):409–420, 1992.
14. N. Echols, D. Milburn, and M. Gerstein. Molmovdb: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res.*, 31:478–482, 2003.
15. M. Gerstein and W. Krebs. A database of macromolecular motions. *Nucleic Acids Res.*, 26:4280–4290, 1998.
16. J. Haile. *Molecular Dynamics Simulation: elementary methods.* Wiley, New York, 1992.
17. B. M. Hespenheide, A. Rader, M. Thorpe, and L. A. Kuhn. Identifying protein folding cores from the evolution of flexible regious during unfolding. *J. Mol. Gra. Model.*, 21:195–207, 2002.
18. D. Jacobs. Generic rigidity in three-dimensional bond-bending networks. *J. Phys. A: Math. Gen.*, 31:6653–6668, 1998.
19. D. Jacobs and M. Thorpe. Generic rigidity percolation: The pebble game. *Phys. Rev. Lett.*, 75(22):4051–4054, 1995.
20. D. J. Jacobs, A. Rader, L. A. Kuhn, and M. Thorpe. Protein flexiblility predictions using graph theory. *Proteins Struct. Funct. Genet.*, 44:150–165, 2001.
21. J. Janin and S. Wodak. Structural domains in proteins and their role in the dynamics of protein function. *Prog. Biophys. Mol. Biol.*, 42:21–78, 1983.
22. P. Karplus and G. Schulz. Prediction of chain flexibility in proteins. *Naturwissencschaften*, 72:212–213, 1985.
23. L. E. Kavraki, P. Svestka, J. C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.
24. O. Keskin, R. Jernigan, and I. Bahar. Proteins with similar architecture exhibit similar large-scale dynamic behavior. *Biophys. J.*, 78:2093–2106, 2000.
25. D. E. Kim, C. Fisher, and D. Baker. A breakdown of symmetry in the folding transition state of protein l. *J. Mol. Biol.*, 298:971–984, 2000.
26. A. Kolinski and J. Skolnick. Monte Carlo simulations of protein folding. *Proteins Struct. Funct. Genet.*, 18(3):338–352, 1994.

27. S. V. Krivov and M. Karplus. Free energy disconnectivity graphs: Application to peptide models. *J. Chem. Phys*, 114(23):10894–10903, 2002.

28. M. Lei, M. I. Zavodszky, L. A. Kuhn, and M. F. Thorpe. Sampling protein conformations and pathways. *J. Comput. Chem.*, 25:1133–1148, 2004.

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

30. M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253:694–698, 1975.

31. R. Li and C. Woodward. The hydrogen exchange core and protein folding. *Protein Sci.*, 8(8):1571–1591, 1999.

32. J. Ma and M. Karplus. The allosteric mechanism of the chaperonin groel: a dynamic analysis. *Proc. Natl. Acad. Sci. USA*, 95:8502–8507, 1998.

33. V. Maiorov and R. Abagyan. A new method for modeling large-scale rearrangements of protein domains. *Proteins*, 27:410–424, 1997.

34. E. L. McCallister, E. Alm, and D. Baker. Critical role of $\beta$-hairpin formation in protein g folding. *Nat. Struct. Biol.*, 7(8):669–673, 2000.

35. V. Muñoz, E. R. Henry, J. Hoferichter, and W. A. Eaton. A statistical mechanical model for $\beta$-hairpin kinetics. *Proc. Natl. Acad. Sci. USA*, 95:5872–5879, 1998.

36. S. Nauli, B. Kuhlman, and D. Baker. Computer-based redesign of a protein folding pathway. *Nature Struct. Biol.*, 8(7):602–605, 2001.

37. W. Nichols, G. Rose, L. T. Eyck, and B. Zimm. Rigid domains in proteins: an algorithmic approach to their identification. *Proteins*, 23:38–48, 1995.

38. S. Prusiner. Prions. *Proc. Natl. Acad. Sci. USA*, 95(23):13363–13383, 1998.

39. A. Rader, B. M. Hespenheide, L. A. Kuhn, and M. Thorpe. Protein unfolding: Rigidity lost. *Proc. Natl. Acad. Sci. USA*, 99(6):3540–3545, 2002.

40. A. Singh, J. Latombe, and D. Brutlag. A motion planning approach to flexible ligand binding. In *7th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, pages 252–261, 1999.

41. G. Song. *A Motion Planning Approach to Protein Folding*. Ph.D. dissertation, Dept. of Computer Science, Texas A&M University, December 2004.

42. G. Song, S. Thomas, K. Dill, J. Scholtz, and N. Amato. A path planning-based study of protein folding with a case study of hairpin formation in protein G and L. In *Proc. Pacific Symposium of Biocomputing (PSB)*, pages 240–251, 2003.

43. M. J. Sternberg. *Protein Structure Prediction*. OIRL Press at Oxford University Press, 1996.

44. 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(8):769–778, 1995.

45. X. Tang, B. Kirkpatrick, S. Thomas, G. Song, and N. M. Amato. Using motion planning to study rna folding kinetics. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 252–261, 2004.

46. W. Whiteley. Some matroids from discrete applied geometry. *Contemp. Math.*, 197:171–311, 1996.

47. D. Xie, S. Thomas, J.-M. Lien, and N. M. Amato. Incremental map generation. Technical Report TR05-006, Parasol Lab, Dept. of Computer Science, Texas A&M University, Sep 2005.

48. D. M. Zuckerman. Simulation of and ensemble of conformational transitions in a united-residue model of calmodulin. *J. Phys. Chem*, 108:5127–5137, 2004.