Modeling Protein Flexibility with Spatial and Energetic Constraints

Yi-Chieh Wu[1], Amarda Shehu[2], Lydia E. Kavraki[2,3]

Abstract

Using principal component analysis, cyclic coordinate descent, and conjugate gradient minimization, physical conformations of HIV-1 protease were produced that captured the opening and closing motions of the flaps, thus modeling the flexibility of the binding site.

1. Introduction

Proteins perform a central role in many cellular functions, from providing structural support to assisting in chemical reactions. As such, understanding cell behavior through the study of molecular interactions remains a fundamental issue in biology. Fischer proposed a lock-and-key model, in which the protein and substrate fit each other in the same way a lock complements a key. However, experimental evidence has shown that a more accurate representation may be found in Koshland's induced-fit model, in which the protein and substrate change conformation to facilitate binding. Despite this support for protein flexibility, most current docking methods consider proteins as rigid structures in order to reduce computational complexity, thus limiting their use in applications such as drug design and discovery. We seek to model this flexibility to provide better representations with which we can analyze protein interactions.

2. Problem Statement

Given a starting structure and the major modes that span the conformational space of a protein, generate a set of conformations to capture the flexibility of the protein. That is, capture the most important motions by starting from the initial structure and following in the direction of the major modes. Our work is performed around the native structure, as analysis fails if we are far from the native.

We focus our analysis on HIV-1 protease, a virus protein that assists in the replication of HIV. Much work has been done in designing drugs to block its active site and thus prevent the virus from replicating. Furthermore, the size of the protease is computationally manageable but large enough to address important problems, and there is abundant data characterizing this protein on which we can validate our results.

3. Previous Work

The most accurate representations of protein motion are obtained through simulation techniques such as molecular dynamics and Monte Carlo. Molecular dynamics uses classical Newtonian equations to compute particle motion while Monte Carlo performs a series of random steps to generate a series of conformations. Both methods are computationally intensive, however, particularly for molecules with many degrees of freedom. For example, molecular dynamics can only capture limited protein motion, on the scale of nanoseconds, making it unsuitable for use in applications such as protein docking, as binding interactions take place on a longer time scale with larger-scale motions.

Rice University, Departments of Electrical and Computer Engineering[1], Computer Science[2], and Bioengineering[3], Houston, TX, 77005.

More recently, robotic kinematics have been applied to the study of proteins, reducing molecules to robotic manipulators, with atoms the equivalent of joints and bonds the equivalent of links. Singh, et. al. [SLB99] and Amato, et. al. [ADS04] both use probabilistic roadmaps to sample possible conformations and generate a collision-free, i.e. low-energy, path from a starting configuration to a given goal.

Other approaches to modeling protein flexibility have included soft receptors, selection of specific degrees of freedom, multiple receptor structures, and collective degrees of freedom [She03]. Furthermore, many techniques have relied on dimensional reduction in hopes of simplifying the problem and thus the computational complexity. The most common method uses principal component analysis (PCA) to identify the major modes of motion of the protein. Moll, et. al. perform an expansive search by generating Gaussian perturbations in major mode space and minimizing. Shehu uses cyclic coordinate descent (CCD) [CD03, Lot04] and minimization to generate conformations [She04], and we draw upon her approach in our algorithm.

4. Description of Algorithm

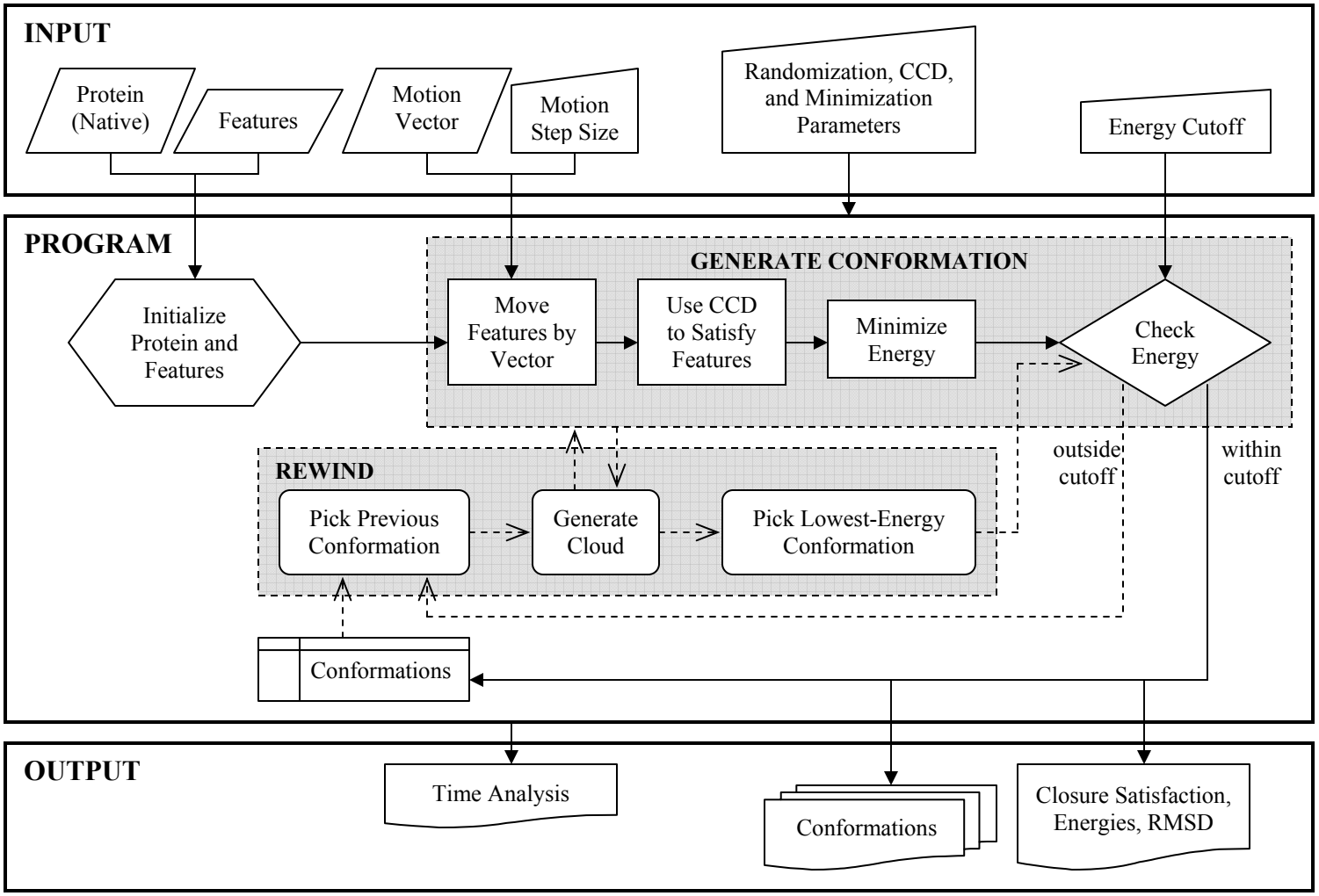

Figure 4.0. Algorithm flowchart.

4.1. Principal Component Analysis and Feature Definition

The most important principal components have direct physical interpretations, representing concerted motion of atom groups. In the case of HIV-1 protease, the first eigenvector corresponds to the opening and closing motion of the flaps, respectively exposing and sealing the binding site. We choose our *features* to be residues along the flaps, which non-coincidentally have the largest displacements (Figure 2). The positions of these features are moved along the PCA at each step to capture flap movement.

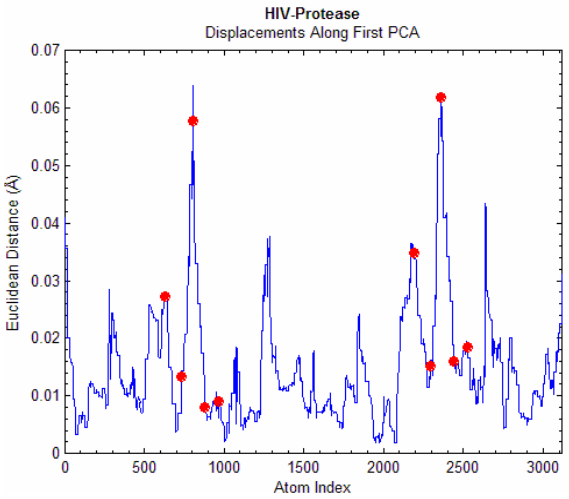

Figure 4.1. Atom displacements along the first PCA. Red circles mark the indices of our chosen features (between residues 41 and 61 and residues 140 and 160).

4.2. Spatial Constraints

Analysis of PDB structures has shown that bond lengths and angles show little variation among structures. Therefore, we assume a rigid geometry model, in which only the bond dihedrals are used as degrees of freedom. This reduces the complexity of our problem and allows us to use cyclic coordinate descent (CCD) [CD03, Lot04] to move features to their constrained positions. CCD is an iterative, heuristic approach to solving inverse kinematics problems and was first developed for use in the field of robotics. It has since been applied successfully in bioinformatics, allowing researchers to close protein loops irresolvable through X-ray crystallography. CCD gradually moves an atom to its constrained location by adjusting one dihedral angle at a time, always trying to minimize to closure distance. Because CCD turns the problem into a series of single-variable minimizations, it is also computationally fast and analytically simple.

We should also note that we use two different versions of CCD, depending on if we wish to satisfy orientation as well. Features along the flaps to be moved along the motion vector use the simpler version of CCD, in which we only attempt to satisfy the carbon-alpha position of the residue. Features at the edges of the flaps we want to keep in its native conformation under the assumption that this will provide the lowest-energy. Thus, we constrain the orientation of these residues as well; that is, we try to satisfy the nitrogen, carbon-alpha, and carbon positions.

4.3. Energetic Constraints

We use a simple probability function to determine energetic feasibility, accepting a conformation if its energy is within 600 kcal/mol of the native energy. Because flap displacement is only valid in a small neighborhood, we also perform a full minimization of the CHARMM energy using conjugate gradient at every step. Since we only define features on the protein backbone, this

minimization also helps fix any sidechains, as they are kept rigid during CCD. Minimization runs the risk that we return to our original structure, so we limit the number of minimizations and assume that the closest conformation of low-energy has the flaps opening or closing.

4.4. Optimization

In order to minimize the time to generate a conformation, we had to balance the time taken closing the loops with CCD and the time taken minimizing energy with conjugate gradient. Increasing the maximum number of steps in CCD forces the rest of the protein to assume a more native conformation, reducing the overall energy and thus reducing the number of iterations we have to apply minimization. However, it also increases runtime linearly as we continue to update dihedral angles. On the other hand, relaxing minimization constraints increases the likelihood that we will accept the final conformation and not have to redo any conformations, but each additional minimization step takes quadratic time in the number of atoms.

To further reduce runtime, we also relaxed the threshold on internal flap features. Unlike the end loop features that serve to keep the rest of the protein in the native conformation, we use these features to model motion and thus only a few CCD steps should be necessary to move these residues to an appropriate position near that specified by the motion vector.

4.5. Backtracking

Moving along the motion vector will ultimately result in non-physical conformations, as the flaps collide with other parts of the protein or are unable to be pushed further while keeping the rest of the protein native. Because we use random path permutations in CCD, a generated conformation for any given step is non-deterministic. Thus, if we encounter a high-energy conformation, rather than trying to move against the same energy barrier, we implemented a simple backtracking technique in which we rewind to a previous conformation and try to push from there, hopefully generating conformations that overcome the previous energy barrier.

5. Experimental Results

We used the structure of 4HVP from the Protein Databank and performed a full minimization to arrive at our native structure. Furthermore, we used only the first major mode in our attempt to model the flap movement of HIV-1 protease, opening and closing the flaps in a way that kept the protein stable.

5.1. Flap Movement

Summary results for our method are provided in tables 5.4, 5.5.1, 5.5.2, and 5.5.3, detailing the movement of the conformation with the highest flap all-atom RMSD for a given run. Because we only updated internal features, movement is concentrated in the flaps while the rest of the protein remains in a native-like conformation. Also, as expected, we were able to generate conformations with higher flap RMSDs when opening the flaps as opposed to closing the flaps, as the former motion is less energetically constrained by the rest of the protein. Whereas Shehu was able to recover energetically feasible conformations with total all-atom RMSDs up to 0.23 Å in the closing direction of the flaps and up to 0.56 Å in the opening direction [She04] using a step size of 0.1, we recovered conformations up to 1.12 Å in the closing direction and up to 1.60 Å in the opening direction. A backbone representation of our recovered conformations is provided in figure 5.1.

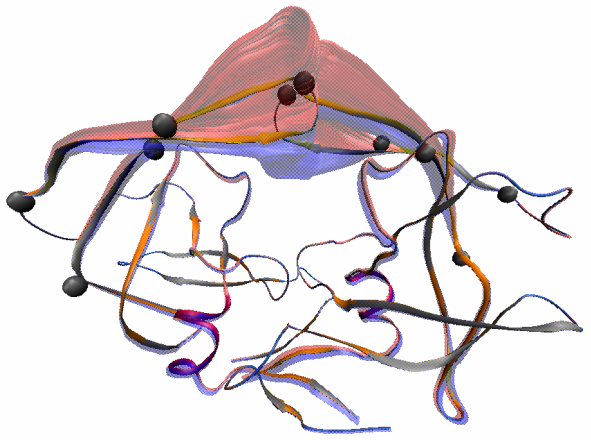

Figure 5.1. Backbone representation of flap movement along the first PCA. Features used are shown as gray spheres.

In contrast to Shehu's results, our energy landscape is non-symmetric, with open-flap conformations tending to have lower energies than close-flap conformations with a similar RMSD. However, we do observe a funnel-like shape, supporting the existence of a well-defined and stable native structure.

5.2. Energy Landscape

Some energy landscapes are provided in figures 5.4.1 and 5.5.1. We observe a funnel-shape, supporting the existence of a well-defined and stable native structure. Open- and close-flap conformations are non-symmetric due to varying energetic constraints, and we see that we are able to capture more conformations around the native.

5.3. Efficiency

On average, it took about thirty seconds to generate a single conformation, with about twelve seconds spent on CCD, about fifteen seconds spent on minimization, and the rest spent on other processes such as energy computation. The time to generate a conformation increased from fifteen seconds for the first conformation to thirty seconds once the number of minimization steps became more constant. If rewinding was necessary, the time to generate the conformation would also be severely higher due to time in randomizing the dihedrals beforehand and in generating the cloud.

5.4. Multiple Runs

Run	Flap All-Atom RMSD (Å)	Flap Backbone RMSD (Å)	Flap Sidechain RMSD (Å)	Rest All-Atom RMSD (Å)	Total All-Atom RMSD (Å)
0	3.125	2.856	3.188	0.483	1.117
1	3.014	2.705	3.086	0.446	1.070
2	2.747	2.492	2.807	0.486	1.007
3	2.966	2.729	3.022	0.488	1.072
4	2.888	2.674	2.939	0.489	1.049

Table 5.4. Results of generating conformations by defining our features to be between residues 41 and 61 and between residues 140 and 160. Multiple runs closing the flaps at a step size of 0.1.

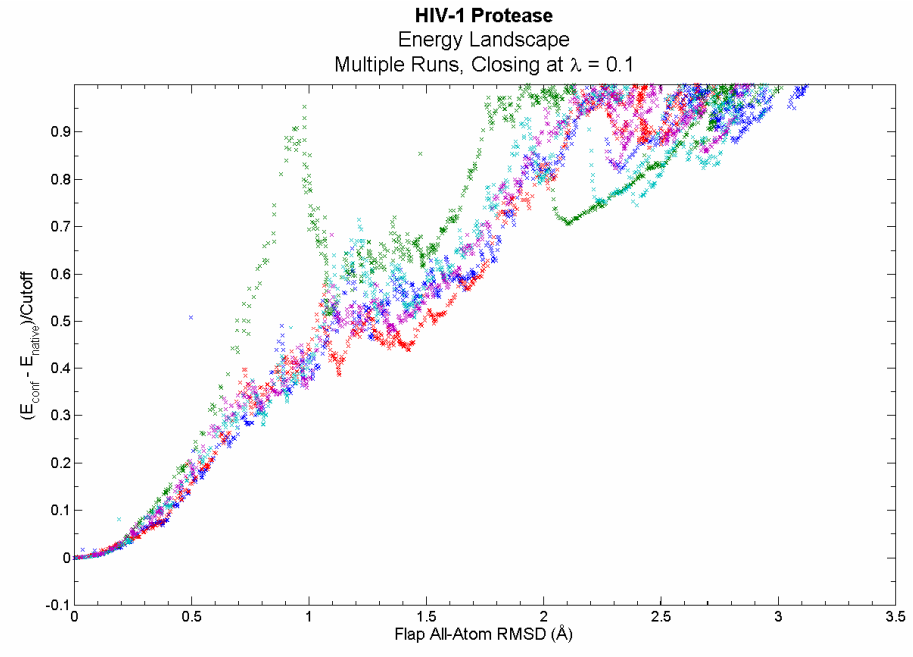

Figure 5.4.1. Energy landscape of HIV-1 Protease. Multiple runs of the same process show roughly the same behavior.

Because our algorithm is non-deterministic due to randomization in CCD paths and in dihedral angles on backtracking, multiple runs of the same process will yield slightly different results. However, the variation in RMSDs is minimal, on the order of what could be attributed to conformational noise.

5.5. Effects of Process Parameters and Backtracking

We provide a short analysis of the effects of different process parameters and the value of backtracking. Randomization deviation, maximum CCD steps, and maximum minimization iterations were also investigated but are not detailed as they affect process efficiency rather than conformation output.

5.5.1. Effects of Step Size

	Step Size	Flap All-Atom RMSD (Å)	Flap Backbone RMSD (Å)	Flap Sidechain RMSD (Å)	Rest All-Atom RMSD (Å)	Total All-Atom RMSD (Å)
Close	0.1	3.125	2.856	3.188	0.483	1.117
	0.25	2.235	2.104	2.266	0.359	0.804
	0.5	2.097	2.032	2.113	0.337	0.755
	1.0	2.289	2.166	2.319	0.298	0.798
	2.5	2.159	1.993	2.198	0.312	0.764
Open	0.1	4.668	4.027	4.814	0.517	1.599
	0.25	3.421	3.111	3.494	0.356	1.166
	0.5	3.340	3.171	3.454	0.375	1.164
	1.0	2.351	2.030	2.424	0.247	0.802
	2.5	1.643	1.434	1.691	0.240	0.582

Table 5.5.1. Results of generating conformations by defining our features to be between residues 41 and 61 and between residues 140 and 160.

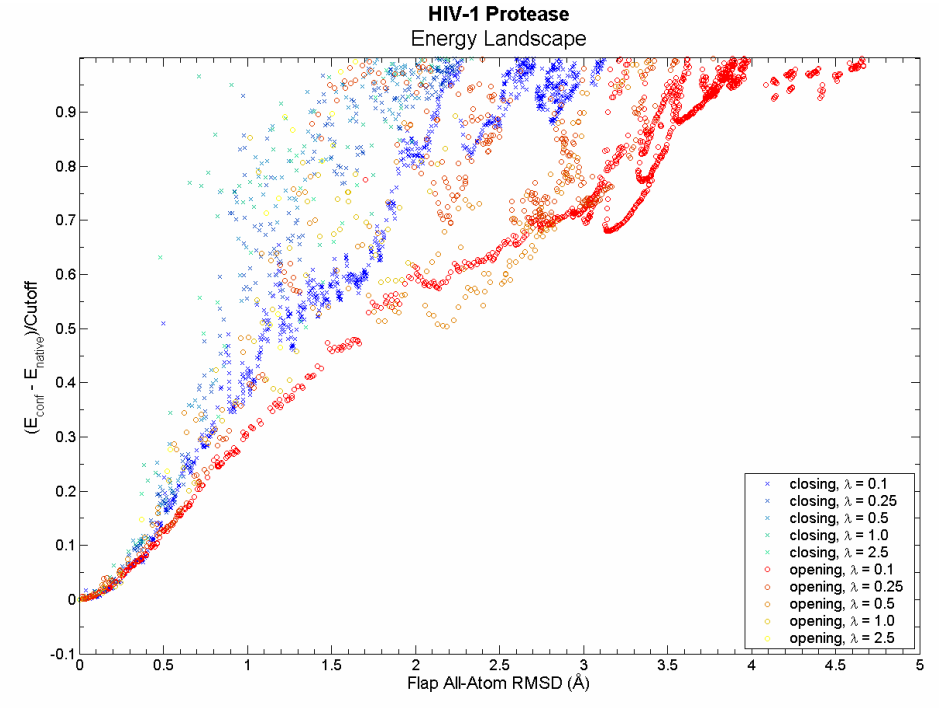

Figure 5.5.1. Energy landscape of HIV-1 Protease. We recover conformations with higher RMSD in the opening direction. We can also push further along the vector is we use a smaller step size.

In general, we are able to push the flaps further along the PCA if we use a smaller step-size. This is not unexpected as closer closure distances cause CCD to make smaller changes to dihedral angles to satisfy spatial constraints. These minor dihedral adjustments are less likely to cause radical atom shifts further down the backbone, resulting in fewer steric clashes. A smaller step-size provides the further advantage of generating more conformations, thus providing a more accurate depiction of the energy landscape. However, more conformations also entail longer processing time to simulate the same flap movement.

5.5.2. Effects of Feature Definition

	Step Size	**Flap All-Atom RMSD (Å)**	**Flap Backbone RMSD (Å)**	**Flap Sidechain RMSD (Å)**	**Rest All-Atom RMSD (Å)**	**Total All-Atom RMSD (Å)**
Close	0.1	2.982	2.892	3.004	0.660	1.156
	2.5	2.551	2.437	2.578	0.551	0.982
Open	0.1	2.353	1.901	2.452	0.479	0.891
	2.5	2.517	2.229	2.584	0.411	0.909

Table 5.5.2. Results of generating conformations by defining our features to be between residues 40 and 60 and between residues 139 and 159.

It is interesting to note the effect of feature definition on our results. Table 5.5.1 agrees with our expectations in that we are able to recover more open-flap conformations as the flaps are less energetically constrained by the rest of the protein. However, table 5.5.2 shows that by moving all features one residue back (though still within the flaps), we are now able to generate more close-flap conformations. This second choice of features may generate conformations stressing the unfavorable interactions between the beta sheets of flaps and the neighboring anti-parallel beta sheets.

5.5.3. Effects of Backtracking

	Step Size	Flap All-Atom RMSD (Å)	Flap Backbone RMSD (Å)	Flap Sidechain RMSD (Å)	Rest All-Atom RMSD (Å)	Total All-Atom RMSD (Å)	Percent Increase
Close	0.1	2.367	1.904	2.468	0.420	0.868	32.0
	0.25	2.179	1.927	2.237	0.370	0.792	2.57
	0.5	2.078	1.972	2.103	0.343	0.751	0.914
	1.0	1.709	1.647	1.724	0.261	0.610	33.9
	2.5	1.152	1.110	1.163	0.230	0.434	87.4
Open	0.1	3.339	2.634	3.492	0.357	1.140	39.8
	0.25	3.139	2.791	3.219	0.328	1.070	8.98
	0.5	2.005	1.685	2.077	0.177	0.675	66.6
	1.0	1.197	1.647	2.015	0.156	0.652	96.4
	2.5	1.134	1.199	1.369	0.175	0.467	44.9

Table 5.5.3. Results of generating conformations by defining our features to be between residues 41 and 61 and between residues 140 and 160. Same experiments as for table 5.5.1 but exits the process when a high-energy conformation is generated. Efficiency of using backtracking is provided in the rightmost column, which gives percent increases in flap all-atom RMSD going from table 5.5.3 (where backtracking is not implemented) to table 5.5.1 (where backtracking is implemented).

Backtracking provided varying gains in output, with increases in flap all-atom RMSD ranging from 0.914 % to 96.4 %. In general, backtracking for larger step sizes only works the first few times, as continual backtracking rewinds to the same range of conformations and encountering the same energy barrier.

Summary

We provide an approach to generating physical conformations of a protein along its most important motions, thereby modeling the flexibility of the binding site. We satisfy spatial constraints using cyclic coordinate descent and take into account energetic constraints using conjugate gradient minimization and a simple backtracking mechanism. Our algorithm has been demonstrated on the first principal component of HIV-1 protease to simulate the opening and closing of the flaps.

Acknowledgements

We would like to thank the members of the Physical and Biological Computing Group at Rice University for their support. This work was supported by the Computer Research Association's Committee on the Status of Women in Computing Research (CRA-W) Distributed Mentor Project.

References

[ADS03] N. M. Amato, and K. A. Dill, and G. Song. (2003). Using Motion Planning to Map Protein Folding Landscapes and Analyze Folding Kinetics of Known Native Structures. Journal of Computational Biology, 10, 239-255.

[CD03] A. A. Canutescu and R. L. Dunbrack. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12: 963-972, 2003.

[Lot04] I. Lotan. (2004). Algorithms exploiting the chain structure of proteins. PhD Thesis, *Stanford University*.

[She03] A. Shehu. (2003). Flexible Receptor- Flexible Ligand Docking. [http://cnx.rice.edu/content/m11456/latest/].

[She04] A. Shehu. (2004). Sampling Biomolecular Conformations with Spatial and Energetic Constraints. MS Thesis, *Rice University*.

[SLB99] A. P. Singh, and J. C. Latombe, and D. L. Brutlag. (1999). A Motion Planning Approach to Flexible Ligand Binding. *Proc. 7th ISMB*, 252-261.