Improved Sampling of Protein Folding Landscapes using Molecular Dynamics

Manasi Vartak[1] mvartak@wpi.edu
Lydia Tapia[2] ltapia@cs.tamu.edu
Shawna Thomas[2] sthomas@cs.tamu.edu
Nancy M. Amato[2] amato@cs.tamu.edu

[1] Worcester Polytechnic Institute, Massachusetts [2] PARASOL Lab, Department of Computer Science, Texas A&M University, Texas
http://parasol.tamu.edu

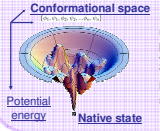
Parasol
Smarter computing.
Texas A&M University

Protein Folding

The function of a protein depends on its three dimensional structure called conformation.
Protein folding problems:

1. Predict the tertiary structure of a protein from its amino acid sequence.

2. Find folding pathways to the known tertiary structure (our work)
   -- Understand the folding process to design better structure prediction methods and to study diseases caused by misfolding e.g. Alzheimer's disease



Folding pathways are studied using potential landscapes of the protein. Different landscapes yield different folding behavior. The lowest point in the landscape is the native, folded state.

Motivation

- **Problem:** Current simulations using the **P**robabilistic **R**oadmap **M**ethod (**PRM**) can sample a limited conformational space (C-space)
- **Cause:** Sampling is based solely on the native state
- **Alternative:** Use other sources of data e.g. Molecular Dynamics (MD) data for sampling.
  -- MD simulations expensive for all but the smallest of proteins.
  -- Overcome limitation by using MD data for small overlapping fragments
- **Benefits:** -- Reduce dependence on the native state
  -- More realistic sampling of C-space with accurate MD data
  -- Augment method with rigidity analysis to get improved sampling.

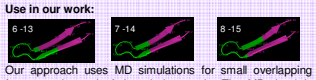
Molecular Dynamics

Molecular dynamics (**MD**) is a versatile tool for studying trajectories of motion for various particles. It consists of the following steps:

- Choosing a force field and potential energy function
- Finding the force on each particle
- Finding and integrating the equations of motion

MD simulations are very accurate and when applied to proteins, can give a single folding pathway. But they are expensive for all but the smallest of proteins.

Use in our work:



Our approach uses MD simulations for small overlapping fragments instead of the whole protein. The MD data* is processed to give a probability distribution for φ and ψ angles of each residues. One distribution is obtained per fragment the residue is present in.

* MD data courtesy of Ken Dill's group at UCSF

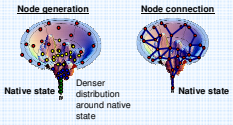
Protein Folding by Probabilistic Roadmap Method

- A roadmap is a graph approximating the potential landscape for the protein. The protein is modeled as an articulated linkage with each amino acid residue having two degrees of freedom (angles φ and ψ).

- A conformation for an *n* residue protein is described by a vector of *2n* angles. Conformational-space (C-space) is the set of all such vectors. The feasibility of a point in this space depends on its potential energy.

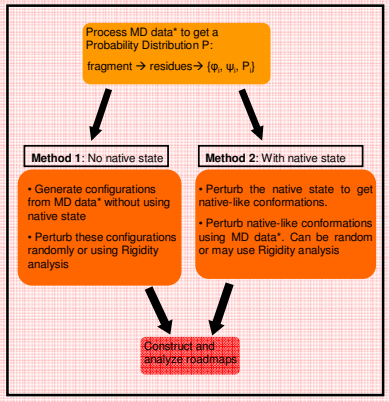
$$\{\phi_1, \psi_1, \phi_2, \psi_2, \ldots \phi_n, \psi_n\}$$

- ***Sampling or node generation***
  Node generation can be biased to some known target conformation. We sample around it, gradually growing out. Conformations are retained based on potential energy. Our present approach uses MD data to sample a larger C-space.
- ***Connecting the roadmap***
  The k (small constant) closest neighbors for each node are connected and an edge weight is assigned reflecting the energetic feasibility of transition.
- ***Extract folding pathways***
  The roadmap produces folding pathways from which secondary structure formation orders can be obtained to validate results.



Our Approach

Our method based on MD data* seeks to improve sampling by reducing dependence on the native state. For this we use data from 2 types of overlapping fragments: 8 residue and 12 residue fragments.

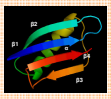


The roadmaps are analyzed for their size, connectivity, secondary structure formation order and generation time. The node distribution is studied using metrics like RMSD, Euclidean distance, potential energy and number of native contacts.

* MD data courtesy of Ken Dill's group at UCSF

Results: Case study of Protein G

Protein G



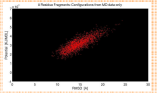
56 residues
(1α helix, 2 β sheets)

Effect of Fragment lengths on landscapes

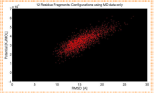
We applied our methods to MD data for 8 and 12 residue fragments of Protein G. 12 residue fragments give better node distribution and lower RMSD – potential limits than those for the 8 residue fragments.

Potential energy vs. RMSD

8 Residue Fragments



12 Residue Fragments

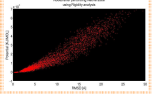


Landscapes from Method 1

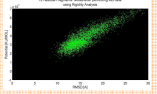
Landscapes from Method 1 (without the native state) have the same general characteristics as the ones generated by existing sampling methods (with the native state).

Potential energy vs. RMSD

Existing Sampling Method

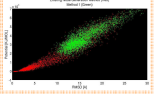


Method 1



Similarity in Landscapes

Existing Sampling Method and Method 1

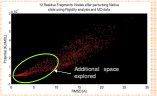


Performance of Method 2

As compared to Method 1, Method 2 samples a larger part of the C-space. Method 2 also produces landscapes very similar to those produced by the existing sampling method. Comparison shows that Method 2 requires fewer nodes and less time to build the roadmap.

Potential energy vs. RMSD

Method 2 compared to Method 1



Performance comparison: Method 2, Existing method

Method	Nodes (N)	Edges (E)	E/N	E+N	Time (min)	SSFO
Existing sampling method	3001	109754	36.57	112755	106.93	3 4 1 2
Method 2	1851	61996	33.49	63847	71.36	3 1 4 2

Conclusion

- 12 residue fragments give better node distribution for Protein G as compared to 8 residue fragments.
- Landscapes produced by Method 1 (without the native state) have the same general trends as those produced from the native state. They also show good clustering.
- Method 2 (with the native state) explores a larger conformational space as compared to Method 1. Landscapes obtained are comparable to ones from the existing sampling method.
- Method 2 requires fewer nodes and less time to generate roadmaps as compared to the existing method.

References

- **Kinetics Analysis Methods for Approximate Folding Landscapes**, L. Tapia, X. Tang, S. Thomas, and N. M. Amato, 15th Int. Conf. on Intelligent Systems for Molecular Biology (ISMB) & 6th European Conf. on Computational Biology (ECCB), to appear, July 2007.
- **Simulating Protein Motions with Rigidity Analysis**, S. Thomas, X. Tang, L. Tapia, and N. M. Amato, in *Proc. of the Int. Conf. on Computational Molecular Biology (RECOMB)*, pp. 394-409, 2006
- **Protein Folding by Motion Planning,** Shawna Thomas, Guang Song, Nancy M. Amato, *Physical Biology*, 2:S148-S155, Nov 2005.
- **Using Motion Planning to Map Protein Folding Landscapes and Analyze Folding Kinetics of Known Native Structures** , Nancy M. Amato, Ken A. Dill, and Guang Song, *J. of Computational Biology (JCB)*, 10(2):239-255, Nov 20022003. Also, in *Proc. of the 6th Int. Conf. on Computational Biology (RECOMB)*, pp.2-11, Apr 2002.

*This research supported in part by NSF grants ACI-9872126, IA-9975018, EIA-0103742, EIA-9805823, ACR-0081510, ACR-0113971, CR-0113974, EIA-0079874, IIS-0118097, by the Texas Higher Education Coordinating Board grant ATP-000512-0261-2001, and by the DOE. **Vartak** supported by the Computing Research Association CRA-W Distributed Mentor Program. **Tapia** supported in part by a NIH Molecular Biophysics Training Grant (T32GM065088) and previously supported by a Department of Education GAANN Fellowship. **Thomas** supported in part by an NSF Graduate Research Fellowship, a PEO Scholarship, a Department of Education GAANN Fellowship, and an IBM TJ Watson PhD Fellowship.