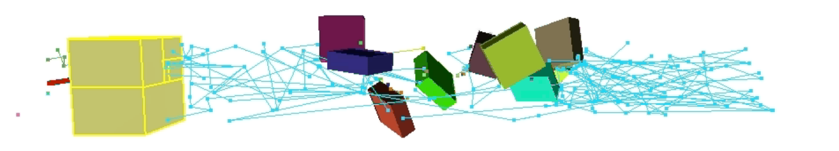
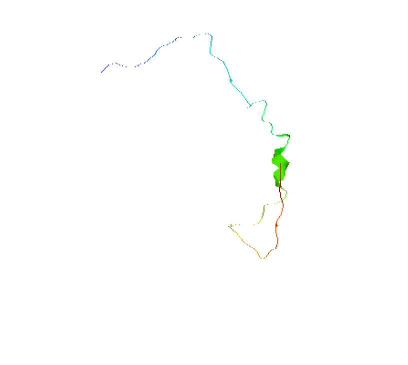




From Robots to Proteins: Randomized Motion Planning for High-Dimensional Problems

Lydia Tapia

Algorithms and Applications Group

Parasol Laboratory

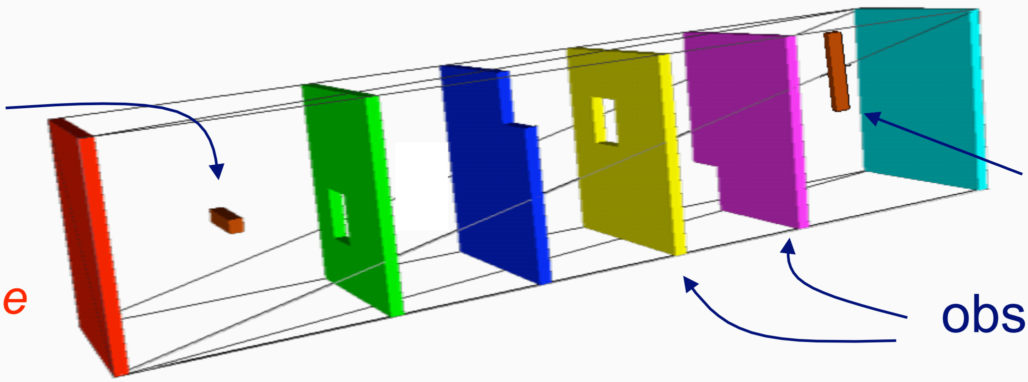
Dept. of Computer Science and Engineering

Texas A&M University

Parasol
Smarter computing.
Texas A&M University

What is motion planning?

- Find a valid path from a start to a goal for a movable object



start

goal

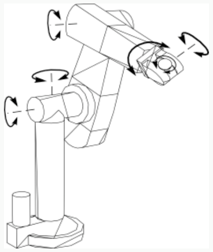
obstacles

valid = collision-free

valid = low energy

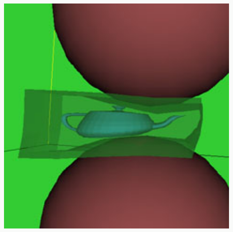
Motions: Robots, Graphics, Molecules

- What do all of these have in common?

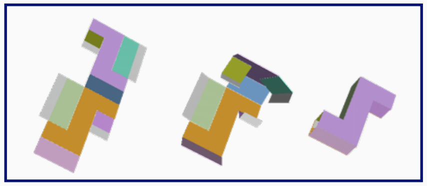


Manipulators

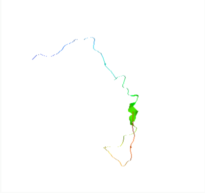
Mobile robots



Deformation

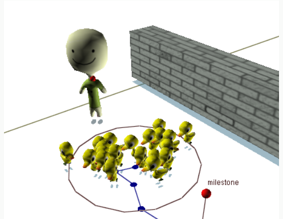


Paper folding

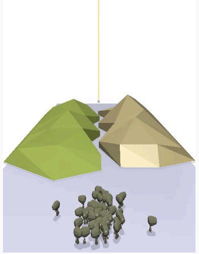


Protein folding

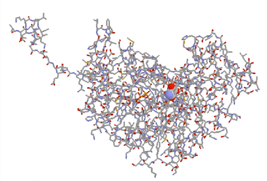
Closed chains



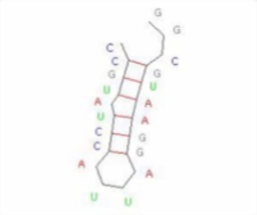
Shepherding



Flocking



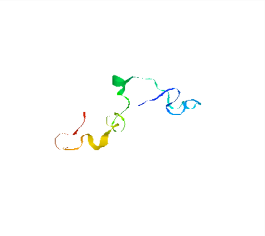
Drug docking



RNA folding

- They are all examples of the motion planning problem
- They can all be solved with the same framework!

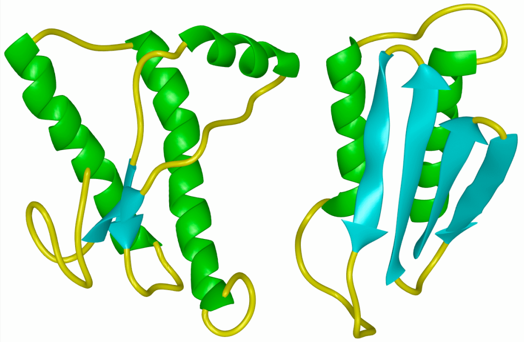
Why Study Folding Pathways?



Importance of Studying Pathways

- Insight into protein interactions & function
  - May lead to better structure prediction algorithms
- Diseases such as Alzheimer's & Mad Cow related to misfolded proteins

prion protein



normal - misfold

Computational Techniques Critical

- Hard to study experimentally (happens too fast)
- Can study folding for thousands of already solved structures
- Help guide/design future experiments

Motion Planning Framework

Robot Abstraction

- How can we develop a single framework to solve all of these different problems?

Point in 3D

●

(x,y,z)

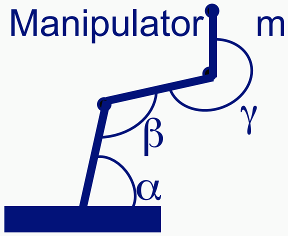
Robot in plane

(x,y,θ)

Rigid body in 3D

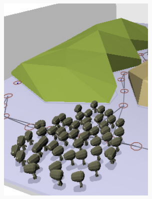
(x,y,z,pitch,roll,yaw)

Manipulator



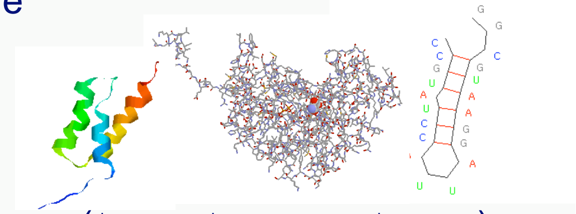
(α,β,γ)

m robots in plane



m*(x,y,θ)

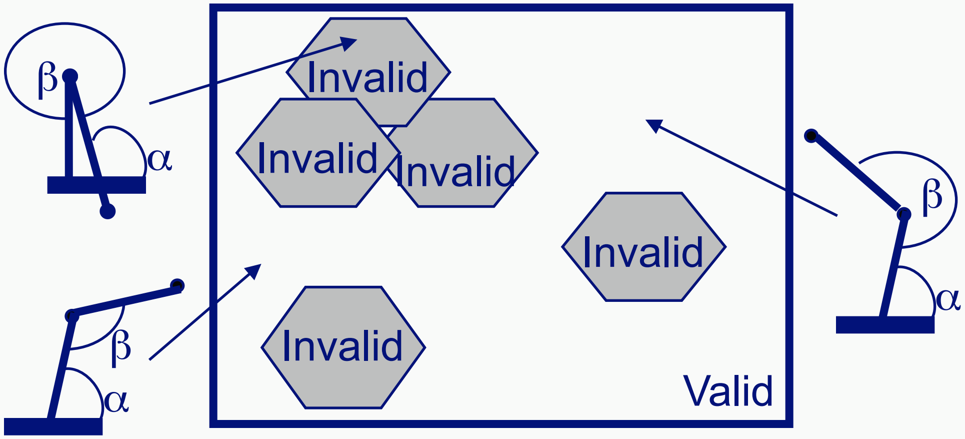
Molecule



$(\phi_1, \psi_1, \phi_2, \psi_2, \dots, \phi_n, \psi_n)$

Configuration Space (C-space):

the set of all object placements



Motion Planning Framework

Probabilistic Roadmap Methods (PRMs)

[Kavraki, Svestka, Latombe, Overmars 1996]

- Idea: Build a model (roadmap) that approximates the topology of the space of Configurations

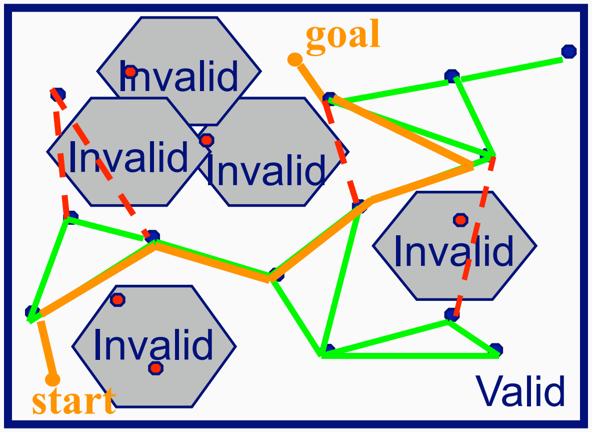
Roadmap Construction

1. Randomly generate robot samples (nodes)
   - discard nodes that are invalid
2. Connect node pairs to form a **roadmap**
   - simple *local planner*
   - discard paths (edges) that are invalid

Query processing

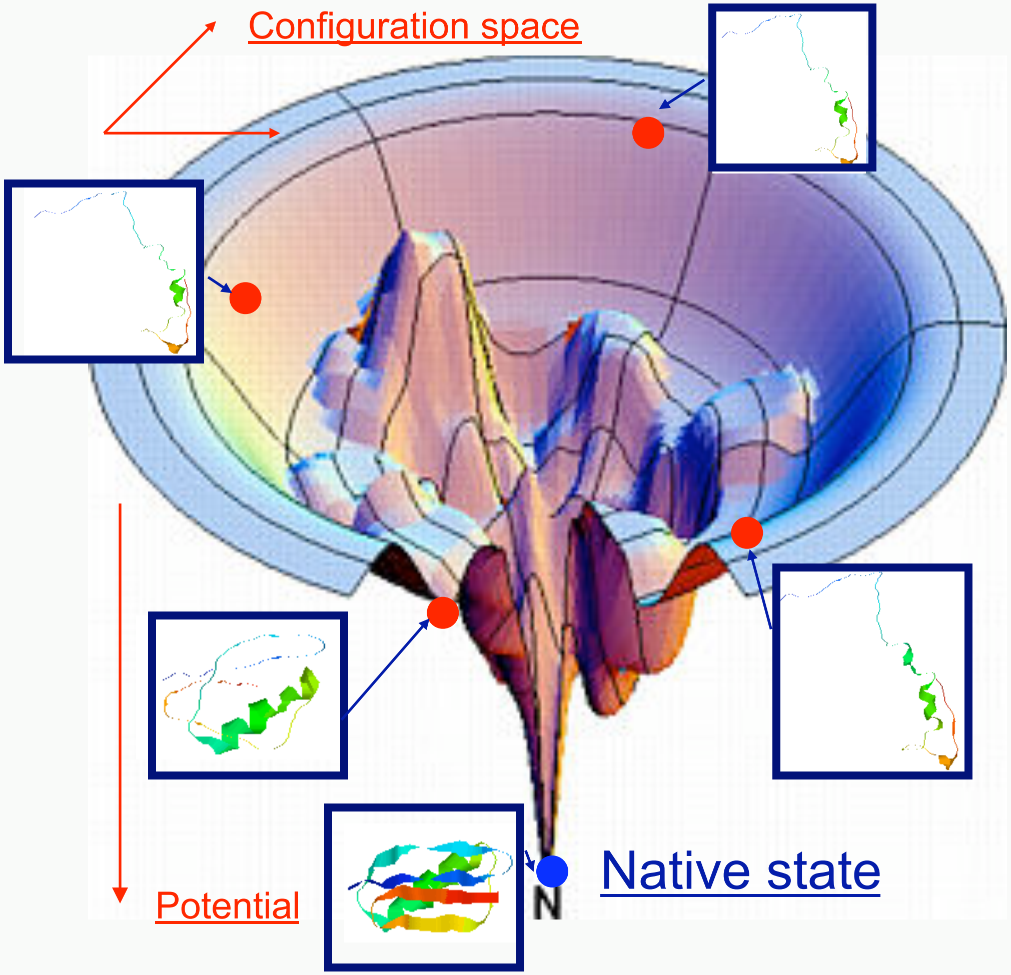
1. Connect *start* and *goal* to roadmap
2. Find path(s) in roadmap between *start* and *goal*

C-space



There's something unique about the space

The Protein Folding Landscape



[Landscapes from Dill and Chan, 1997]

Potential Energy Landscape

- Funnel shape
- Native state is global minimum
- Different proteins ⇔ Different landscapes ⇔ Different folding behaviors

Goal: Build a model (roadmap) of the energy landscape

- Characterize main features
- Extract folding pathways
- Extract folding kinetics

The energy landscape is huge!

Related Work

Simulating Folding & Kinetics

Parasol

	Approach	Folding Landscape	# Paths Produced	Path Quality	Compute Time	Folding Kinetics
Trajectory based	Molecular Dynamics [Levitt 83; Haile 92; Daggett, Levitt 93; Duan & Kollman, 98; Shirts & Pande 00, Boczko & Brooks 95]	No	1	Good	Long	Yes
	Monte Carlo Simulation [Covell 92; Kolinski, Skolnick 94]	No	1	Good	Long	Yes
Statistics based	Master Equation Calculation [Cieplak et al. 98, Ozkan et al. 01, 02, Weikl and Dill 03; Weikl et al. 04]	Yes (required)	N/A	N/A	Fast	Yes
	Statistical Models [Muñoz et.al. 98; Alm, Baker 99; Muñoz, Eaton 99; Baker 00; Matysiak, Clementi 04;Das et al.05]	Yes	0	N/A	Fast	Average
Graph based	SRS and P_{fold} [Apaydin et al. 01, Chiang et al. 06]	Yes	Many	Coarse	Fast	Yes
	Our Roadmap-Based [Song, Amato ICRA 01, JCB 01; Amato et al. JCB 02, Thomas, Tang, **Tapia**, Amato JCB 07, **Tapia**, Tang, Thomas, Amato Bioinformatics 07, Thomas, **Tapia**, AmatoTR08-004, **Tapia**, Thomas, Amato TR08-005; **Tapia**, Thomas, Amato CIS 09]	Yes	Many	Approx. (tunable)	Fast	Yes

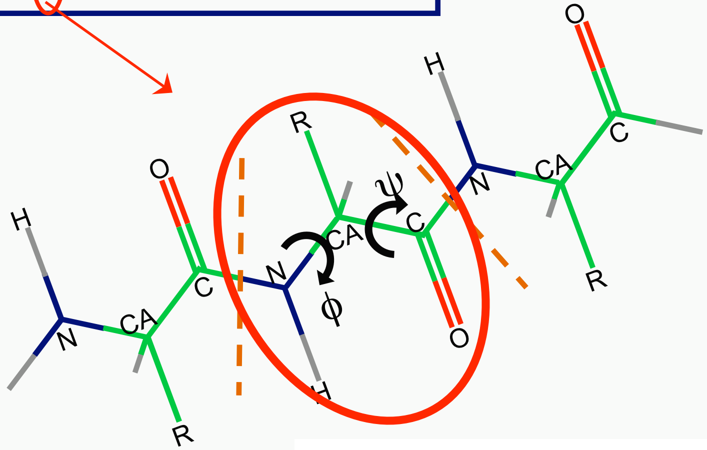
- Other Roadmap-based approaches for studying molecular motions
  - Ligand binding [Singh, Latombe, Brutlag ISMB 99; Bayazit, Song, Amato ICRA 01]
  - RNA Folding [Tang, Kirkpatrick, Thomas, Song, Amato JCB 05; Tang, Thomas, **Tapia**, Amato RECOMB 07; Tang, Thomas, **Tapia**, Giedroc, Amato JMB 08]

Preliminaries:

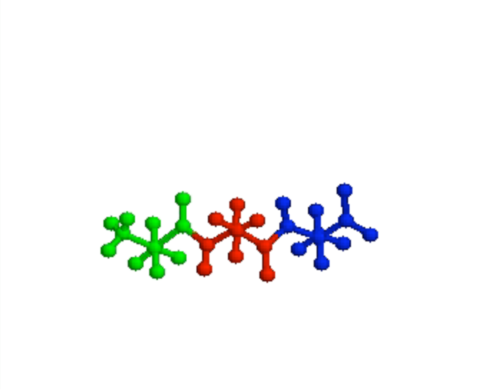
Protein Structure/Model

A protein is a sequence of amino acids/residues, each with 2 torsional degrees of freedom

TTCCPSIVARSNFNVCRLPGTPEALCATYTGCIIIPGATCPGDYAN



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



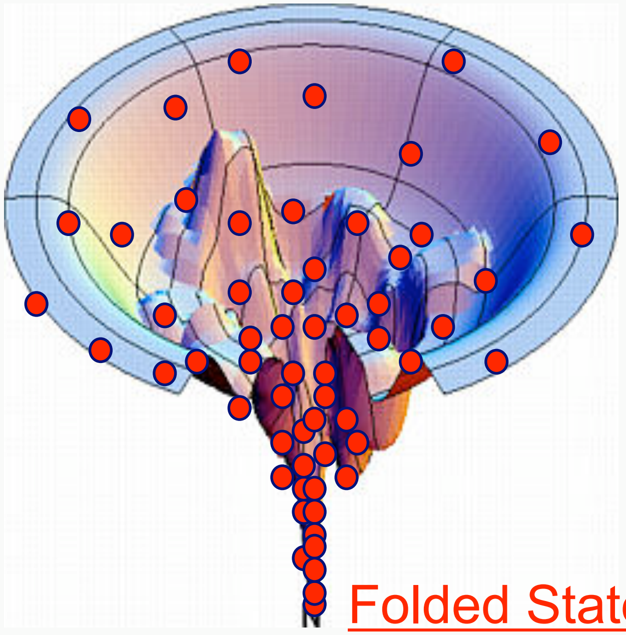
Protein Folding by Motion Planning

Node Generation

- Sample using known target state
- Criterion for accepting a node:

  Compute potential energy *E* of each node and retain it with probability:

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



Denser distribution around target state

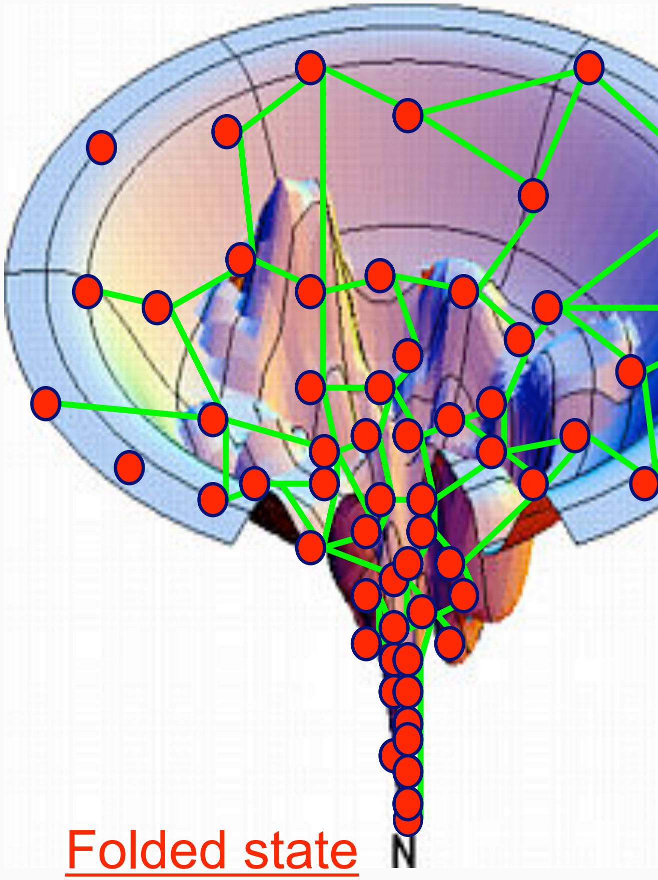
Biased sampling to reduce search space

Our coarse energy function is similar to [Levitt 83] and includes van der Waals, hydrogen bonds, and hydrophobic interaction components

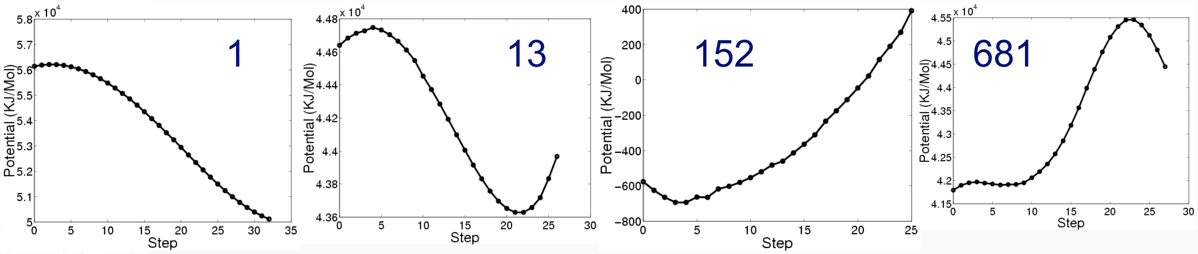
[ICRA'01; RECOMB '01, '06; JCB '02, '07]

Protein Folding by Motion Planning

Node Connection



1. Find k closest nodes for each roadmap node
   - Conformation space distance metric
   - Euclidean, RMSD, Rigidity-Based,…
2. Assign edge weight ***w*** to reflect energetic feasibility



lower weight ⇔ more feasible

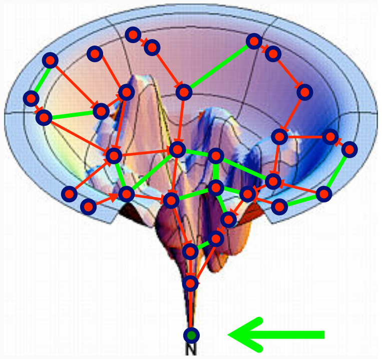
[ICRA'01; RECOMB '01, '06; JCB '02, '07]

Protein Folding

Path Extraction and Analysis

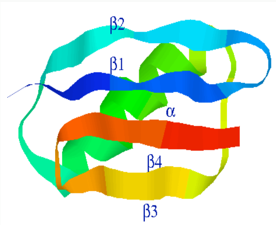
- Roadmap contains **thousands** of folding pathways from unfolded to folded
  - Extract using Dijkstra’s shortest path alg.
  - Analyze pathway’s energy profile, secondary structure formation order, etc.

unfolded

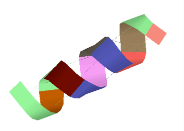


folded

- We **group pathways** based on their secondary structure formation order



Q: Which forms first?



α helix

β sheet

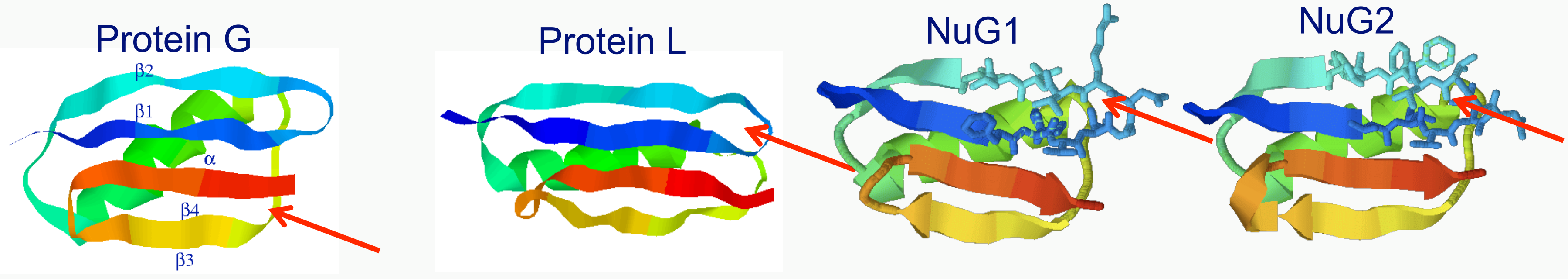
Secondary structure piece is **formed** when it contains most of the native contacts / it is mostly rigid

Do our pathways produce the same orders as seen experimentally?

Protein Folding

Formation Order of G, L, and Mutants

- Proteins G, L, and two mutants of G (NuG1 and NuG2) have similar structure but fold differently [Li, Woodward 99] [Nauli, et al., 01]



Protein G

Protein L

NuG1

NuG2

α, β3β4, β1β2, β1β4

α, β1β2, β3β4, β1β4

Protein	Experimental Order	Roadmap Order	%
G	[α, β1, β3, β4], β2[1] [α, β4], [β1, β2, β3][2]	α, β3-4, β1-2 β3-4, α, β1-2	99.4 0.6
L	[α, β1, β2, β4], β3[1] [α, β1], [β2, β3, β4][2]	β1-2, α, β3-4	100.0
NuG1	β1-2, β3-4[3]	α, β1-2, β3-4	97.6
NuG2	β1-2, β3-4[3]	α, β1-2, β3-4	96.6

[1] Hydrogen out-exchange experiments [Li, Woodward 99]

[2] Pulsed labeling/competition experiments [Li, Woodward 99]

[3] F-value analysis [Nauli, et al., 01]

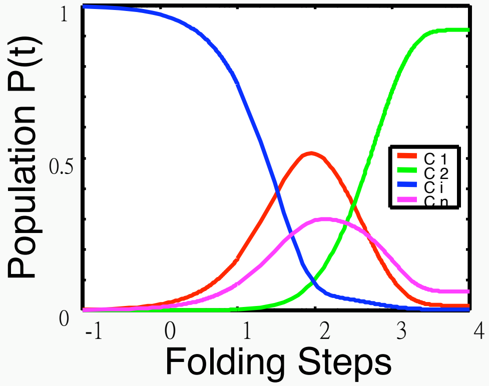
Folding behavior for all four proteins predicted [Thomas, Tang, **Tapia**, Amato JCB 07]

Folding rates for G, NuG1, NuG2 are drastically different [Nauli, et al., 01]

Protein Folding Kinetics

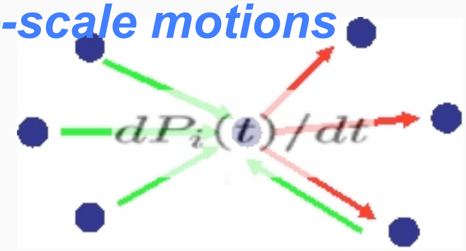
Kinetics is the study of reaction rates

- Folding rates – Faster vs. Slower
- Population kinetics – Change in Conformers
- Validation with Other Experimental Techniques
  - Tryptophan Fluorescence
  - Circular Dichroism
  - H/D Exchange

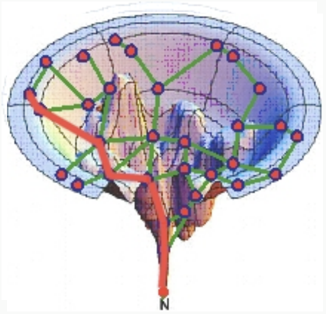


Map-Based Analysis Techniques

Uses local transition probabilities to identify likely large-scale motions



Technique 1: Map-Based Master Equation Calculation (MME)

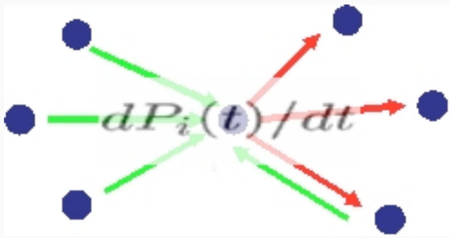


Technique 2: Map-Based Monte Carlo (MMC)

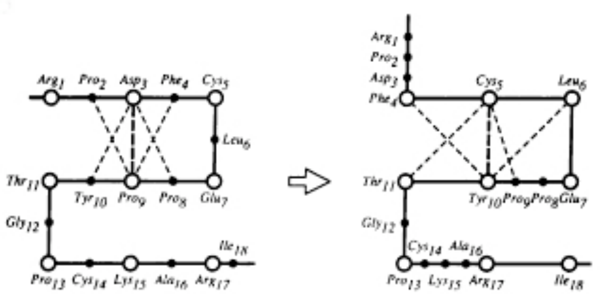
These techniques provide results that can be validated against lab experiment!

Map-Based Technique 1

Map-Based Master Equation (MME)

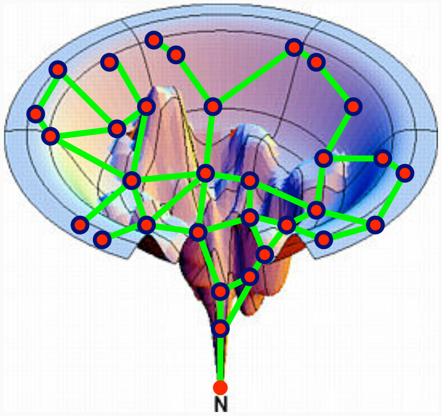


- Master Equation (ME) is a differential equation describing the probability of a process to be in a given state
- Challenge:
  - Usually applied to a detailed model of the energy landscape (lattice, etc.)
  - Thus, limited to small proteins



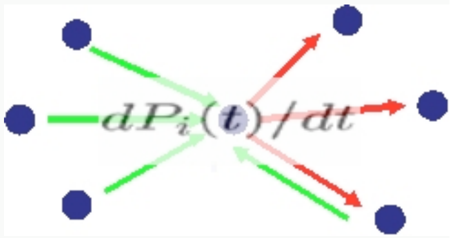
[Hinds and Levitt, PNAS 1992]

- Our solution:
  - Apply to our roadmap (approximate landscape model) instead
  - Roadmap gives model (conformations and transitions) for master equation

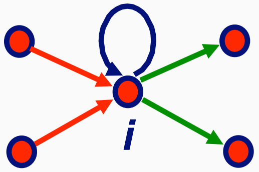


Map-Based Technique 1

Map-Based Master Equation (MME)



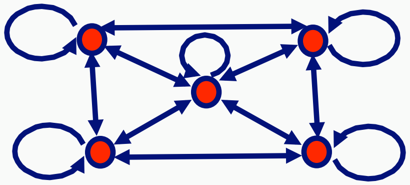
- For conformation *i*, its population over time can be described by:



$$dP_i(t)/dt = \sum_{i \neq j}^{n} (\underline{k_{ji}P_j(t)} - \underline{k_{ij}P_i(t)})$$

*k*ij is a transition probability calculated from edge ij in our roadmap

- The master equation describes the population kinetics of all conformations



$$d\mathbf{p}(t)/dt = M\mathbf{p}(t)$$

$$\begin{cases} M_{ij} = \underline{k_{ji}} & i \neq j \\ M_{ii} = -\underline{\sum_{i \neq j} k_{ij}} & \end{cases}$$

- The solution encodes folding rates (eigenvalues) and important conformation distributions (eigenvectors)

$$P_i(t) = \sum_k \sum_j N_{ik} e^{\lambda_k t} N_{kj}^{-1} P_j(0)$$

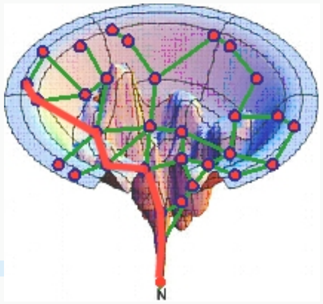
N_{i0} = Boltzmann equilibrium distribution

λ_1 = folding rate (for 2-state folders)

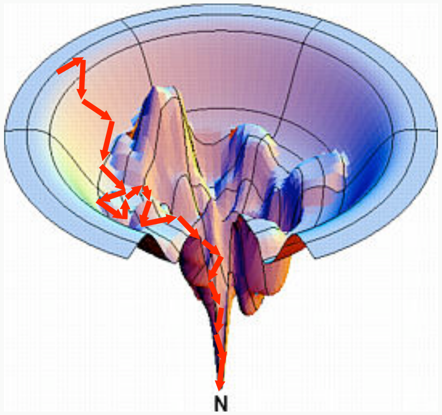
[Kampen 92; Weikl, Plassini, Dill 04]

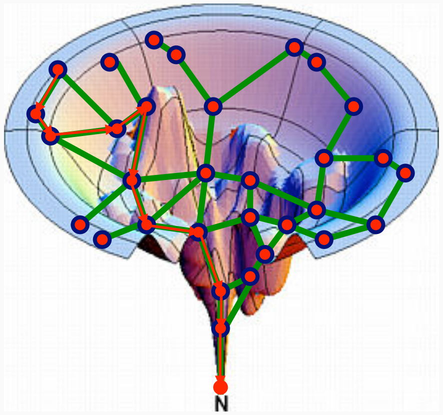
Map-Based Technique 2

Map-Based Monte Carlo (MMC)



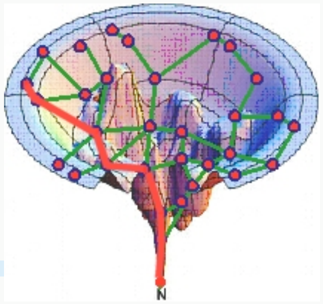
- Monte Carlo (MC) simulation is a random walk on the energy landscape
- Challenge: [Covell, 1992; Kolinski and Skolnick, 1994]
  - At every timestep, MC computes the complete local landscape
  - Limited to small proteins
- Our solution:
  - Apply to our roadmap (approximate landscape model) instead
  - Calculate structure formation from MMC paths

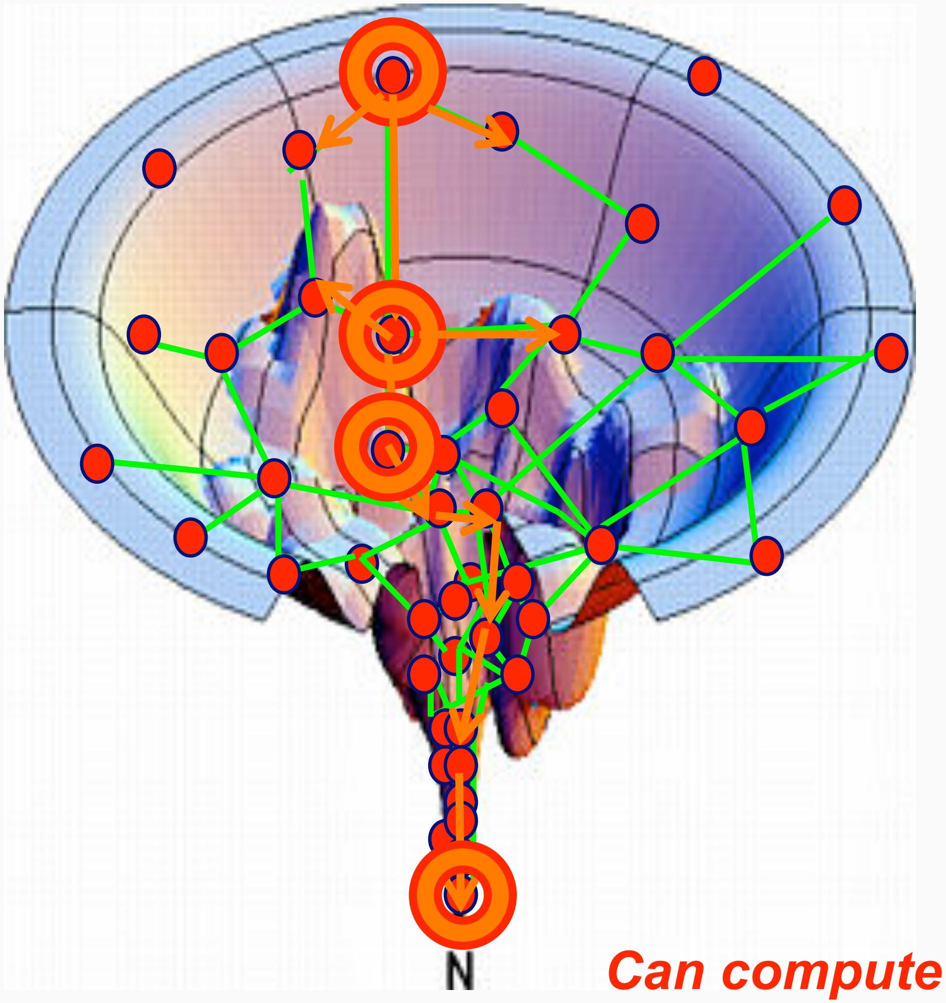




P_{ij} proportional to $1/w_{ij}$

MMC Algorithm





- Start at random unfolded state, current node
- Repeat until maximum number of steps
  - Identify adjacent nodes (neighbors) of current node in the map
  - Calculate the transition probabilities from the edge weight
  - Move to a neighbor probabilistically

Can compute population kinetics and structural features of each conformation in each timestep

Kinetic Case Study

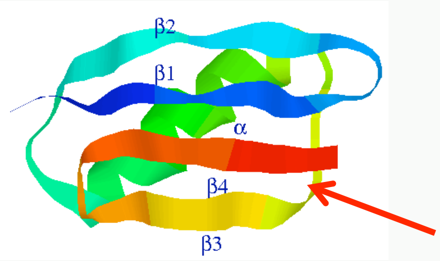
Protein G, NuG1, and NuG2

- Protein G and its mutants NuG1 and NuG2
  - Small, two-state folders
  - G was mutated to alter the hairpin formation order
  - Both have the same secondary and tertiary structure

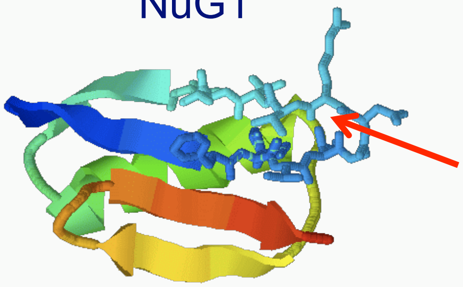
- Our roadmaps captured the secondary structure formation order for Protein G and variants NuG1 and NuG2

[Thomas, Tang, **Tapia**, Amato JCB 07]

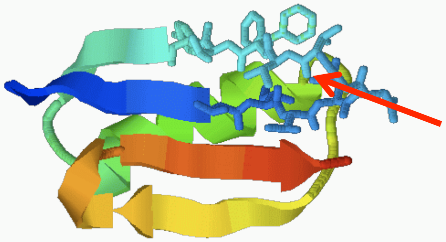
Protein G



NuG1

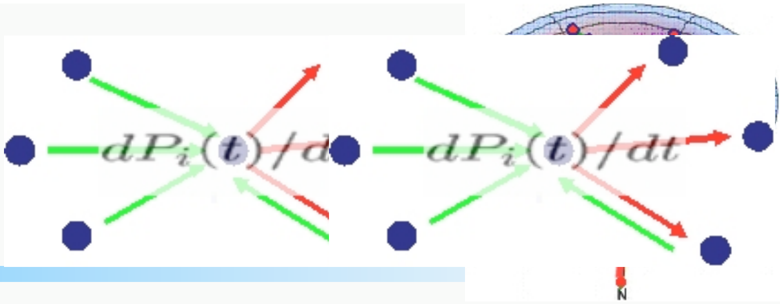


NuG2

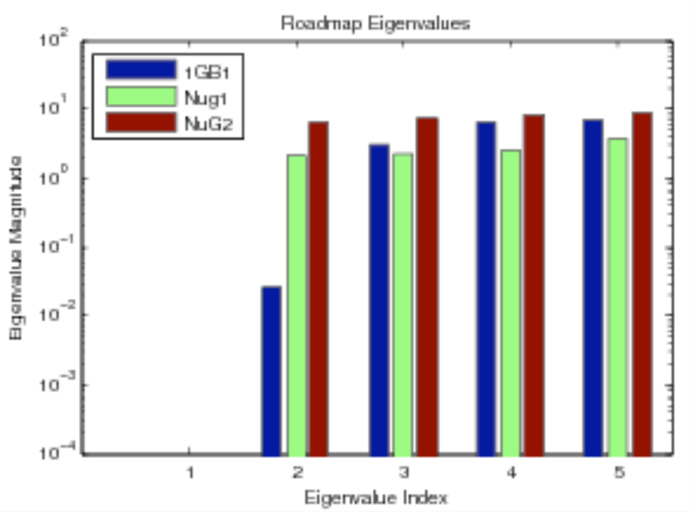


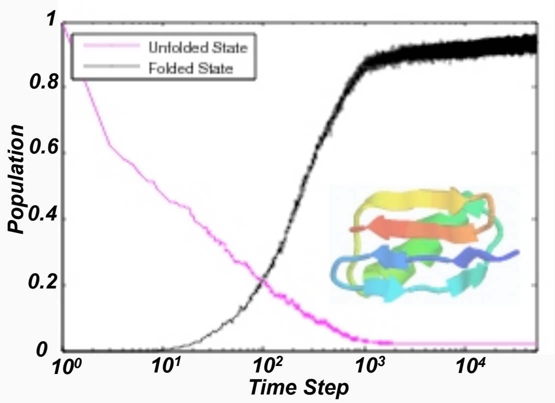
Mutants NuG1 and NuG2 fold 100 times faster than protein G [Nauli et al., 01]

Relative Rates of G, NuG1 and NuG2

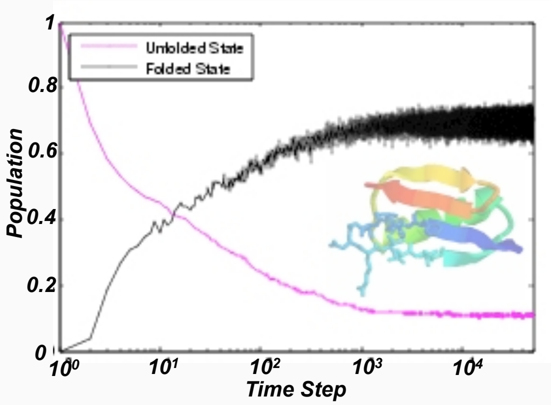


- MME: NuG1 and NuG2 faster than Protein G
- MMC: Faster folding rate of NuG1 and NuG2 also seen in population kinetics

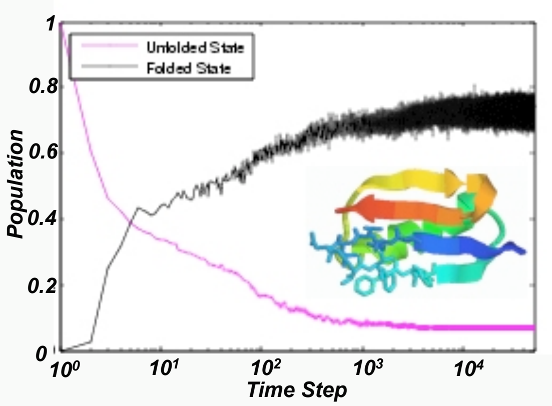




Protein G



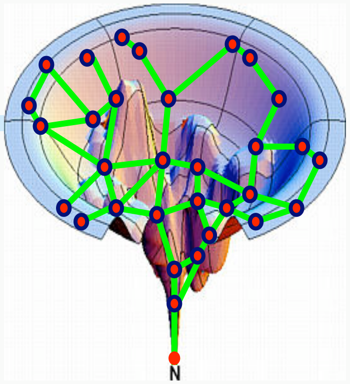
NuG1



NuG2

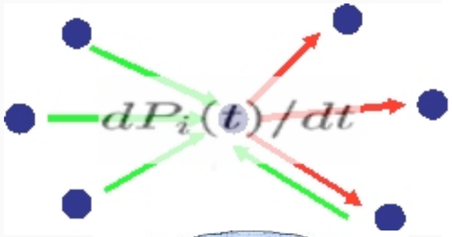
Summary

Map-based Protein Folding Techniques



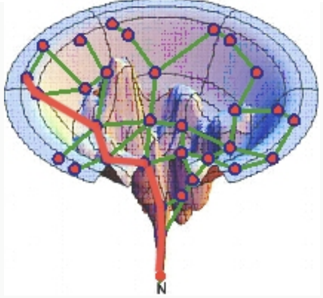
Probabilistic Roadmap Methods for studying protein motions

Uses local transition probabilities to identify likely large-scale motions



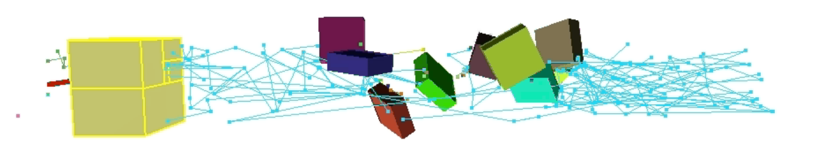
Technique 1: Map-Based Master Equation Calculation (MME)

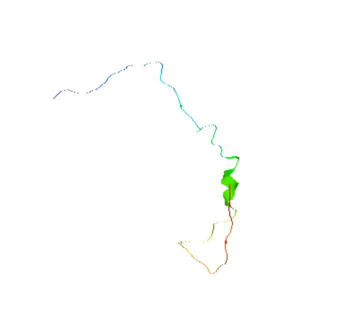
Technique 2: Map-Based Monte Carlo (MMC)



Ability to study time-based structural events

Ability to study a wide-range of structures and folding behaviors





From Robots to Proteins…

Lydia Tapia
Algorithms & Applications Group, Parasol Lab,
Department of Computer Science and Engineering, Texas A&M University
www.parasol.tamu.edu/~ltapia
ltapia@cs.tamu.edu

Many more results at: http://parasol.tamu.edu/groups/amatogroup/foldingserver/

Undergraduate Researchers:

Surbhi Chaudhry, Robotics
Terra Horton, Robotics
Luke Hunter, Protein Folding
Kokil Jadika, Robotics
Kasia Leyk, Protein Folding
Lakshmi Reddy, Robotics
Annette Stowasser, Protein Folding
Manasi Vartak, Protein Folding

Collaborators:

Prof. Nancy M. Amato, Texas A&M
Prof. David Giedroc, Biochemistry, Indiana Univ.
Prof. J Martin Scholtz, Medical and Cellular Medicine, Texas A&M
Bryan Boyd, Texas A&M
Prof. Marco Morales, ITAM
Roger Pearce, Texas A&M
Sam Rodriguez, Texas A&M
Xinyu Tang, Google
Shawna Thomas, Texas A&M

Acknowledgements:

Dr. Mauricio Lasagna,Texas A&M Univ.
Dr. Mark Moll, Rice University
Prof. A.J. Rader, Indiana University-Purdue University