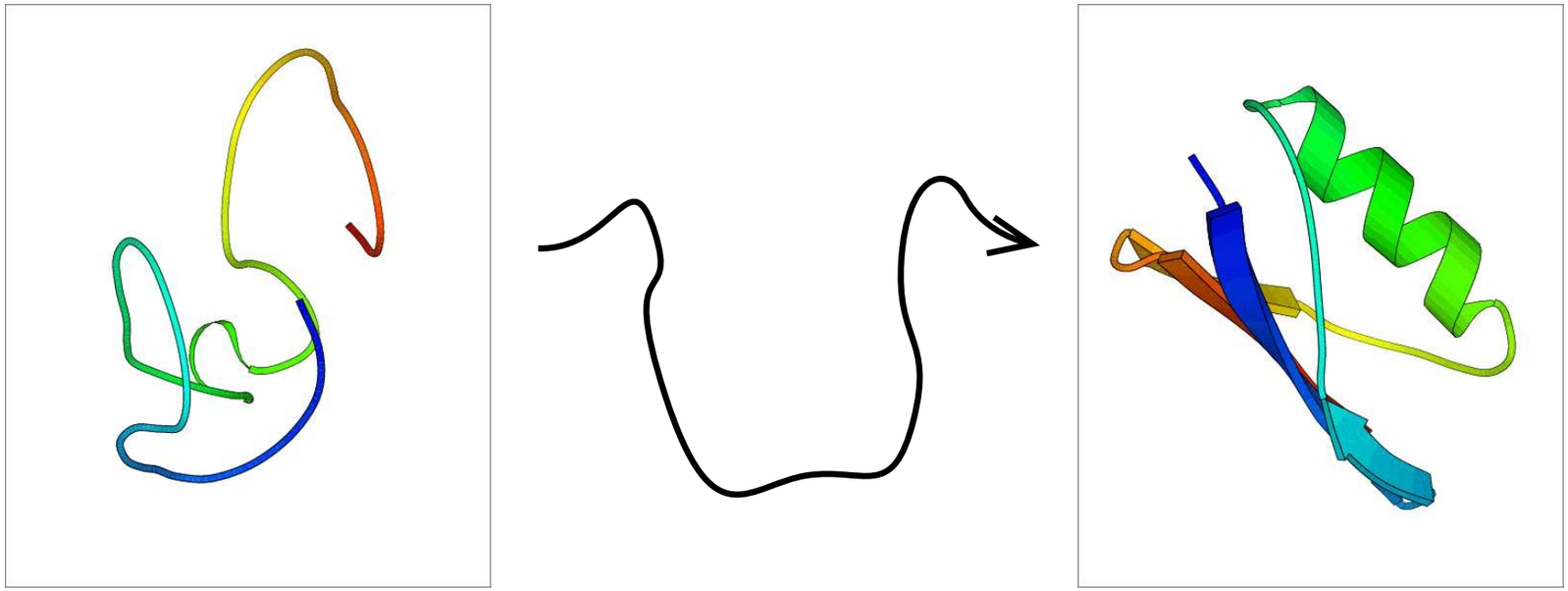
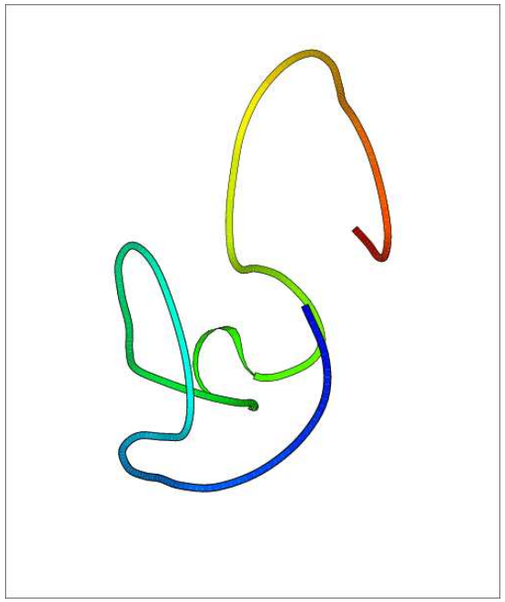
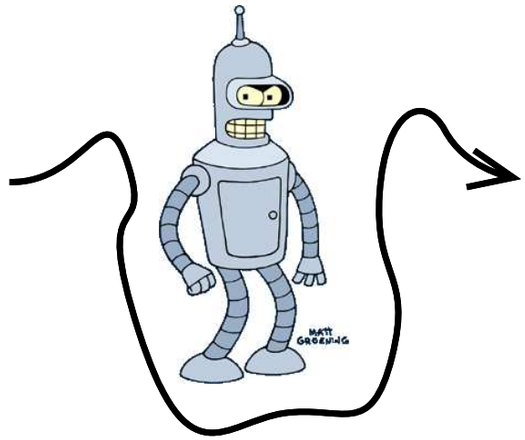
Protein Folding by Robotics

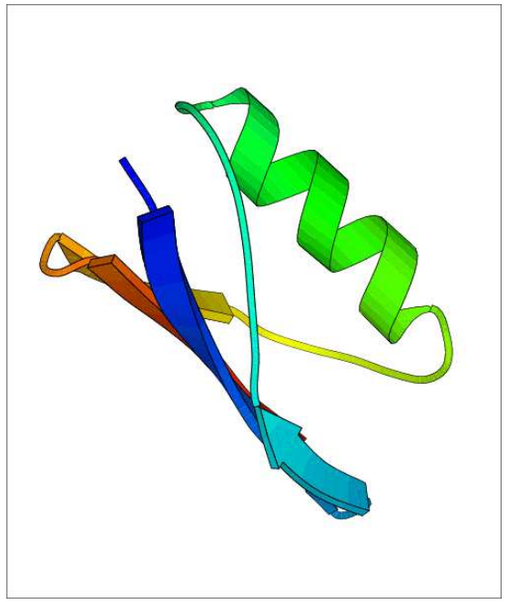
1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion



Protein Folding by Robotics

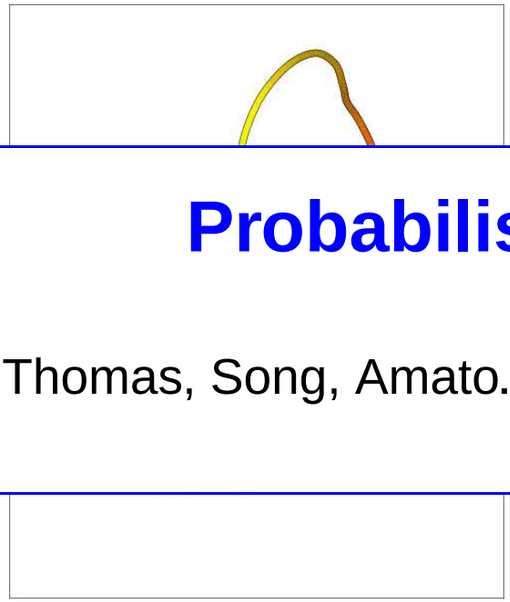


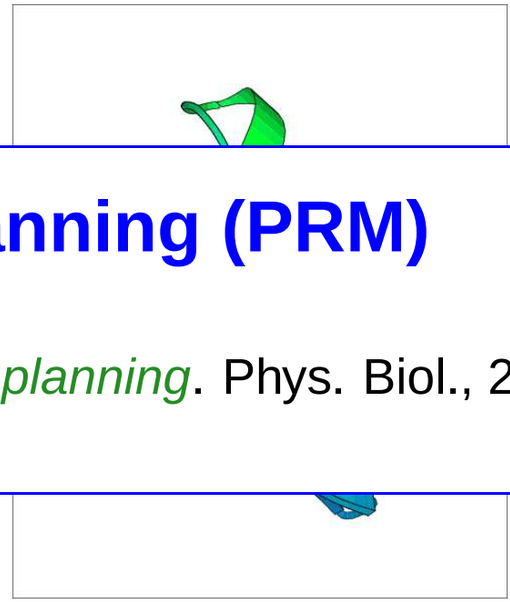




Protein Folding by Robotics

1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion



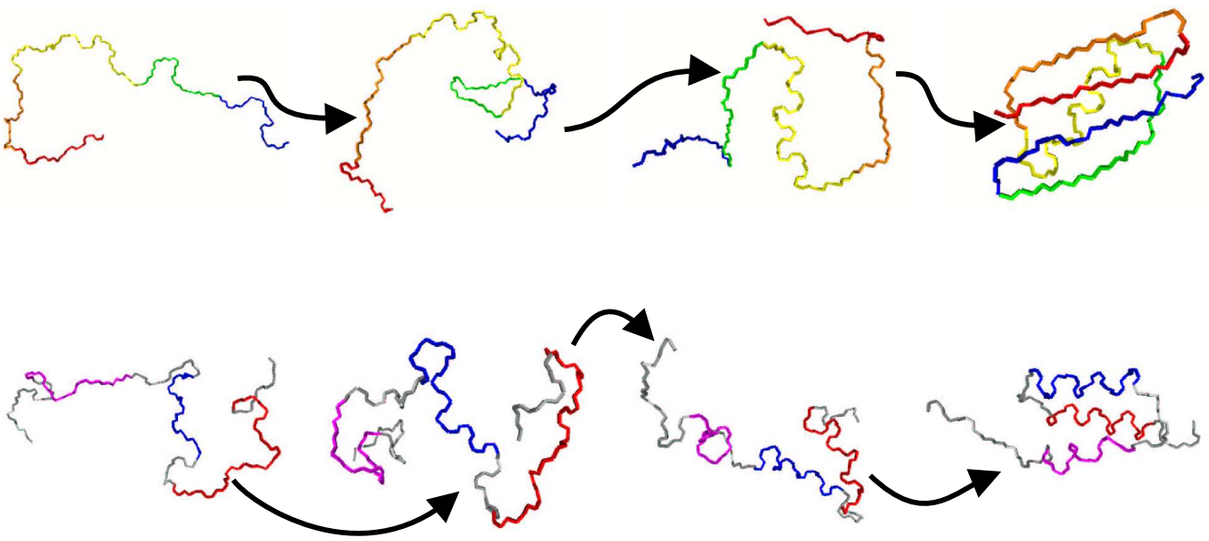


Probabilistic Roadmap Planning (PRM)

[Thomas, Song, Amato. *Protein folding by motion planning*. Phys. Biol., 2005]

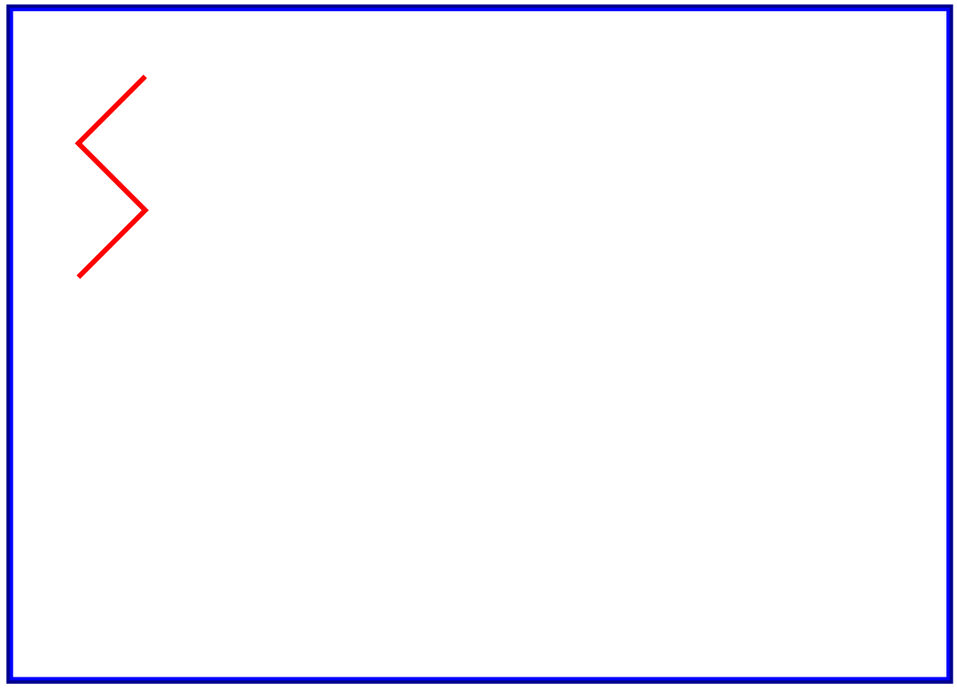
Aims

- Find good quality folding paths (into given native structure)
  - no structure prediction!
- Predict formation orders (of secondary structure)



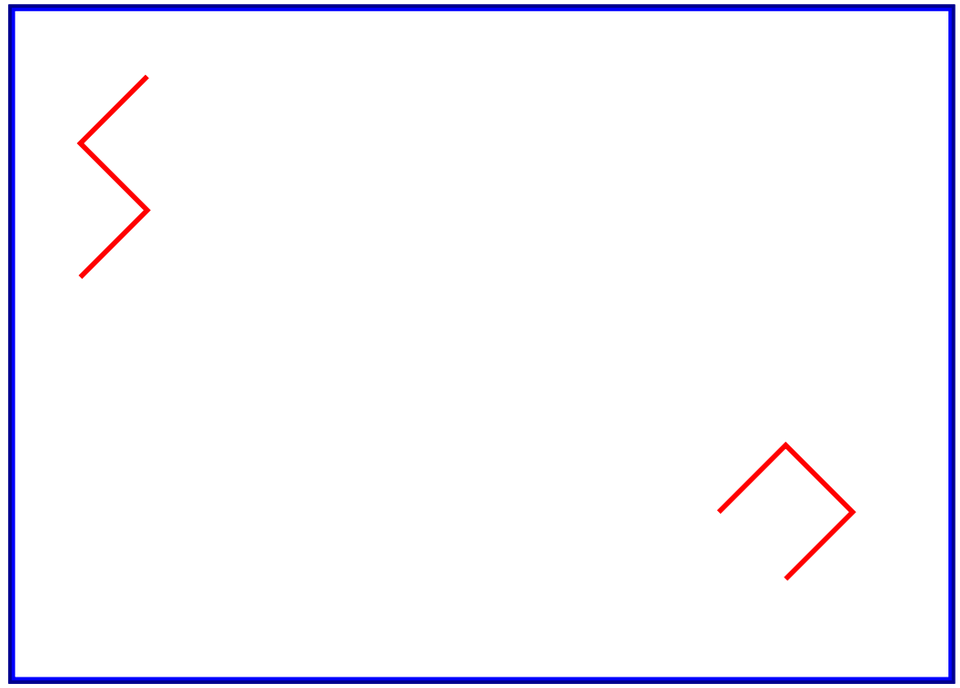
Motion planning

- Motion planning



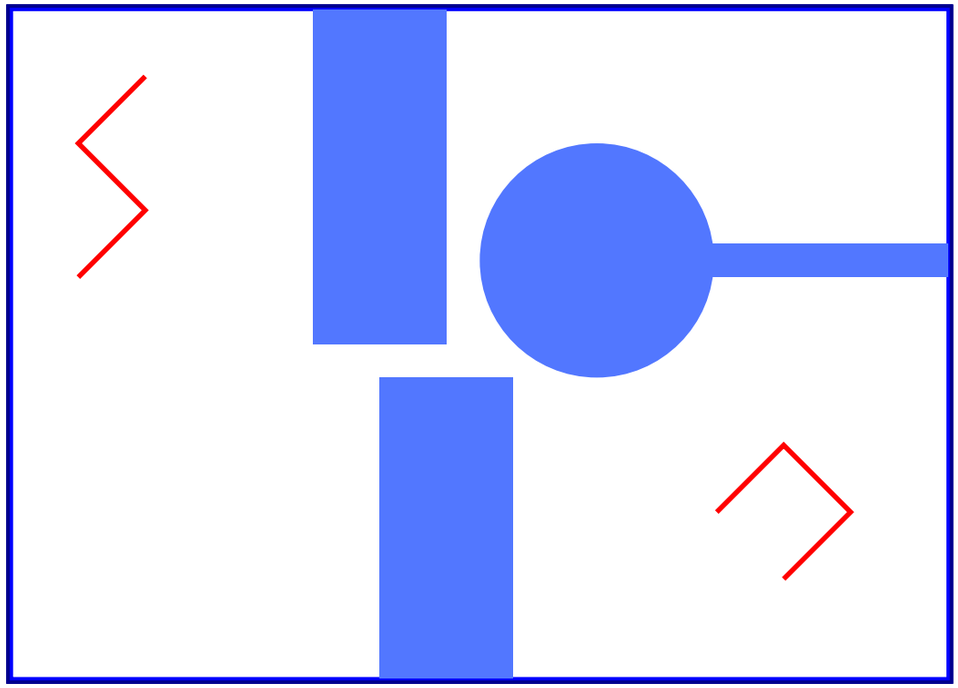
Motion planning

- Motion planning



Motion planning

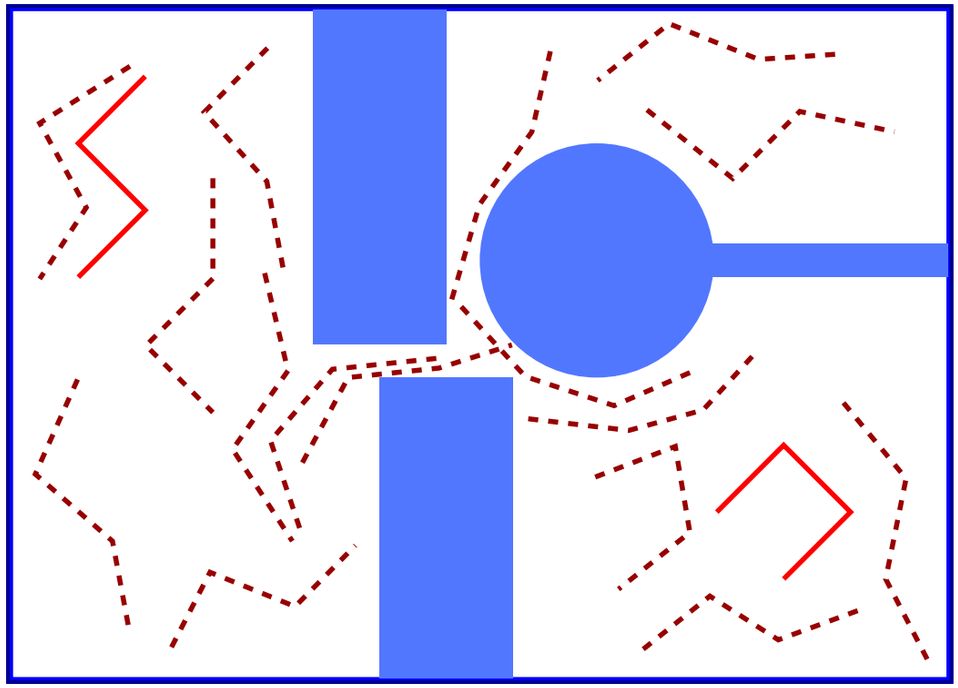
- Motion planning



- Probabilistic roadmap planing

Motion planning

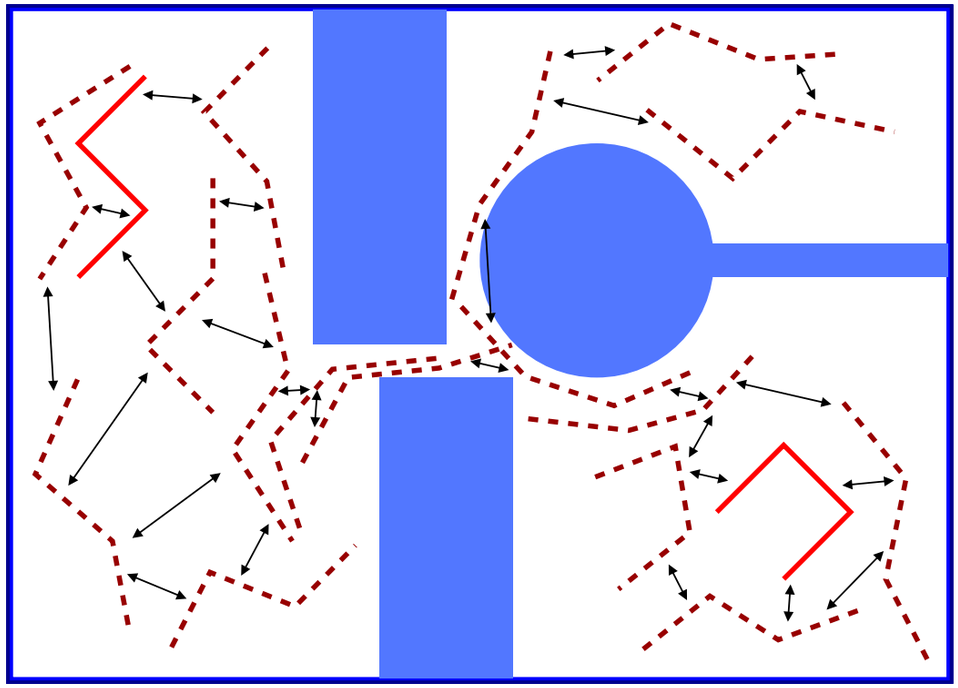
- Motion planning



- Probabilistic roadmap planing
  - Sampling of configuration space Q

Motion planning

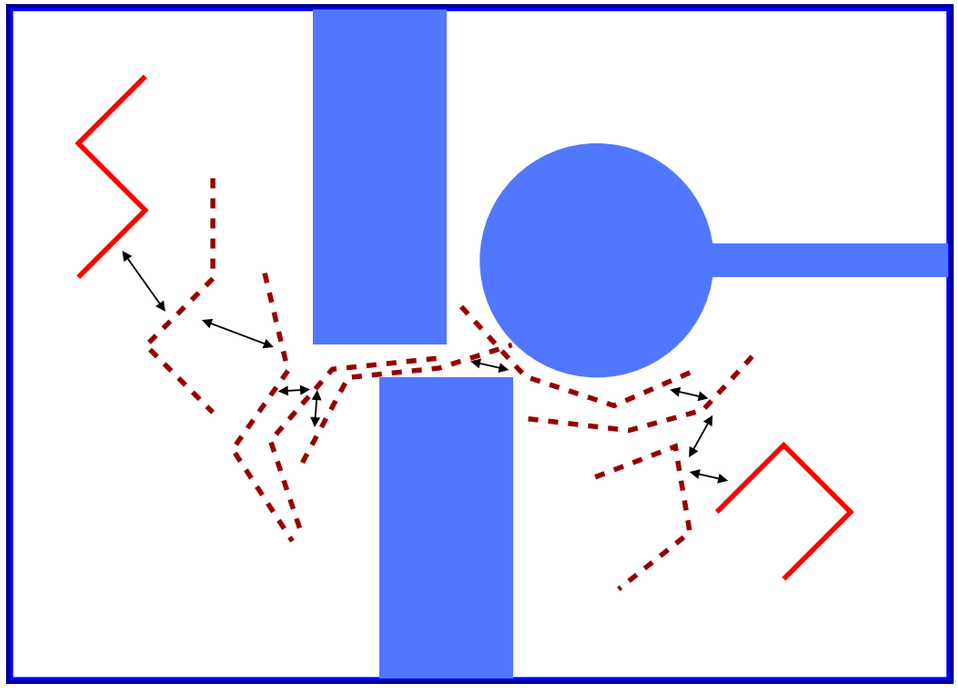
- Motion planning



- Probabilistic roadmap planing
  - Sampling of configuration space Q
  - Connecting nearest configurations by a (simple) *local planner*

Motion planning

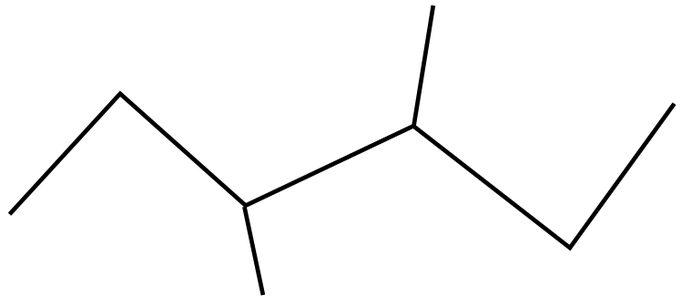
- Motion planning



- Probabilistic roadmap planing
  - Sampling of configuration space Q
  - Connecting nearest configurations by a (simple) *local planner*
  - Apply graph algorithms to “roadmap”: Find shortest path

More on PRM for motion planning

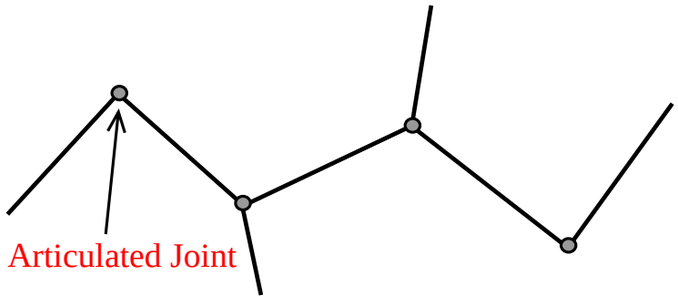
- tree-like robots



More on PRM for motion planning

1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

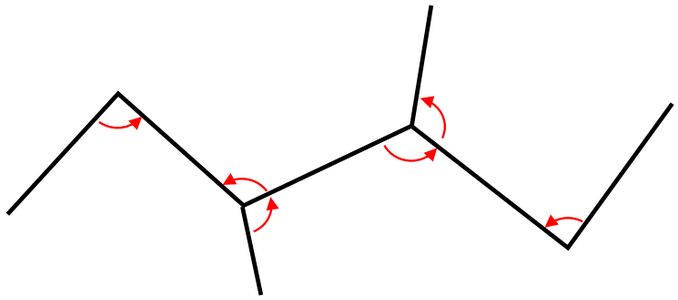
- tree-like robots (*articulated robots*)



More on PRM for motion planning

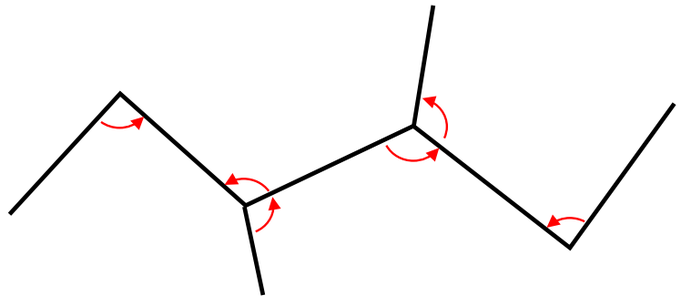
1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

- tree-like robots (*articulated robots*)



More on PRM for motion planning

- tree-like robots (*articulated robots*)



- confi guration = vector of angles
- confi guration space

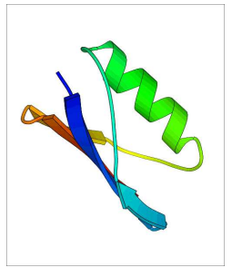
$$Q = \{q \mid q \in S^n\}$$

  - S — set of angles
  - n — number of angles = degrees of freedom (dof)

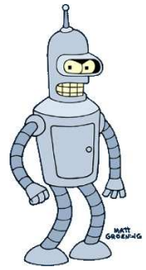
Proteins are Robots (aren't they?)

1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

- Obvious similarity

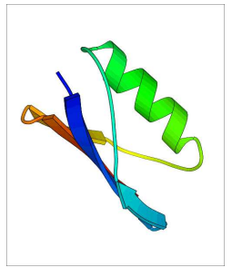


==

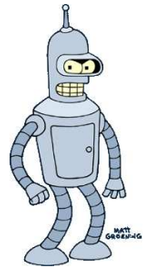


Proteins are Robots (aren't they?)

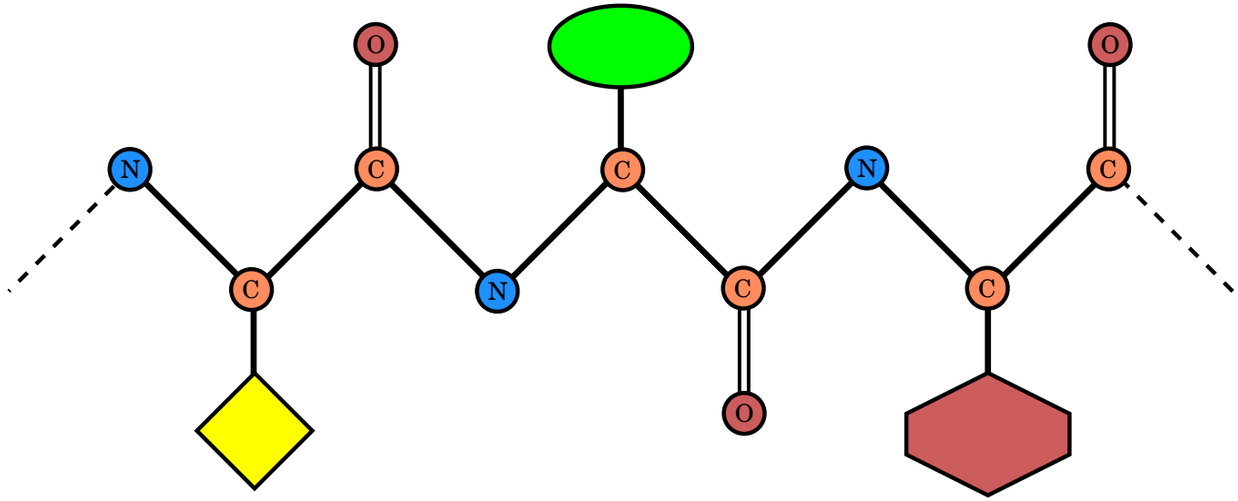
- Obvious similarity ;-)



==

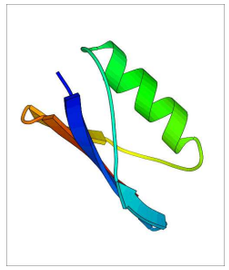


- Our model

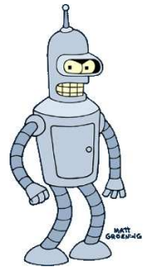


Proteins are Robots (aren't they?)

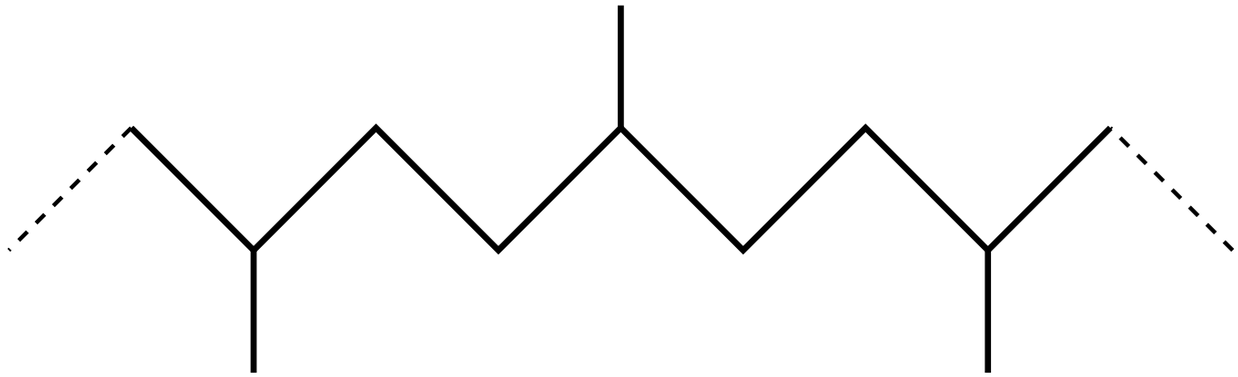
- Obvious similarity ;-)



==

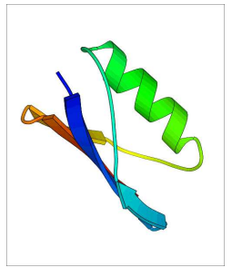
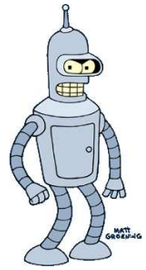


- Our model

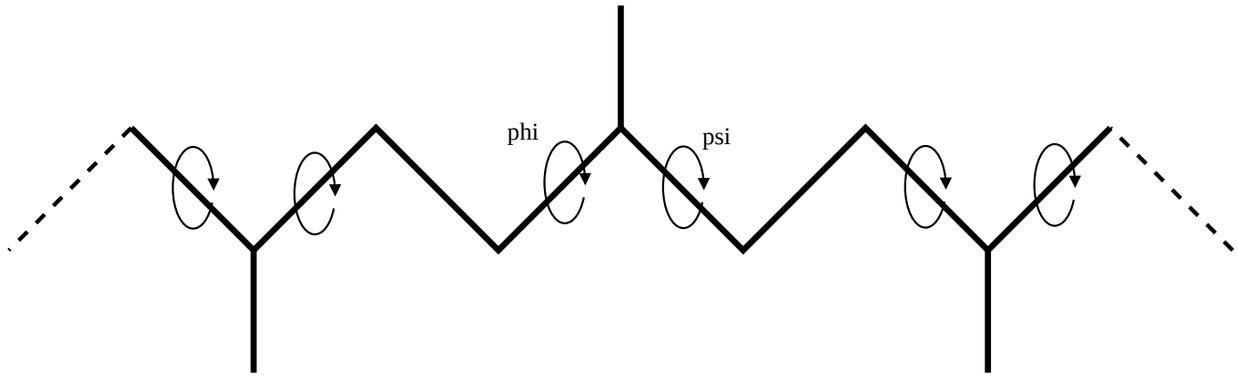


Proteins are Robots (aren't they?)

- Obvious similarity ;-)

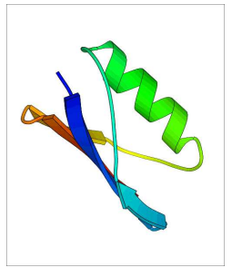

==


- Our model

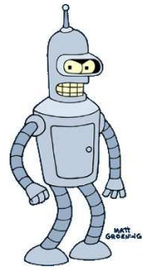


Proteins are Robots (aren't they?)

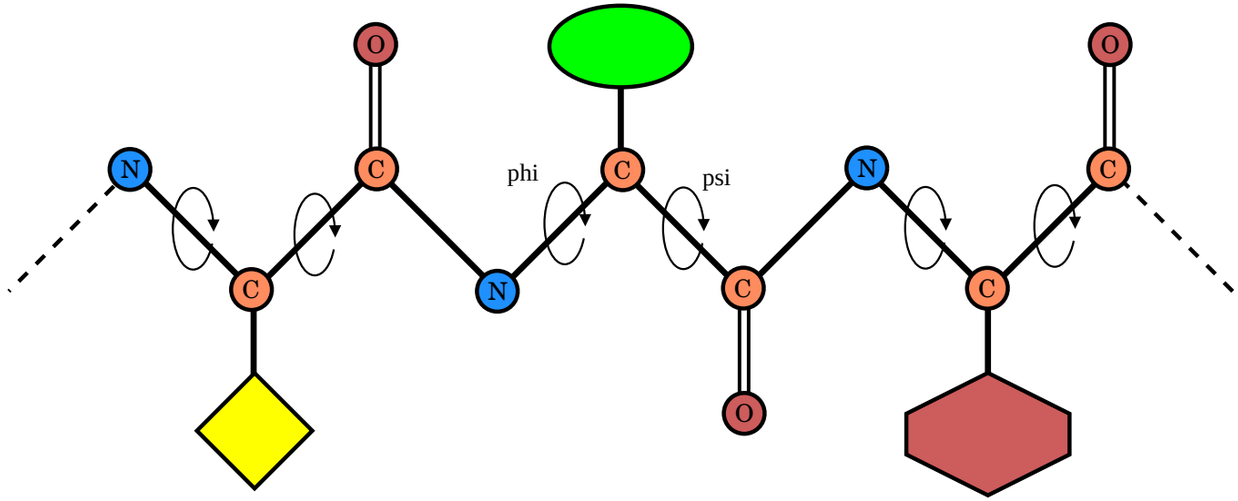
- Obvious similarity ;-)



/==

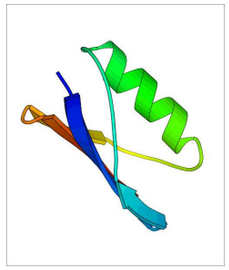


- Our model

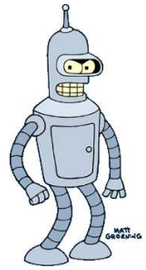


Proteins are Robots (aren't they?)

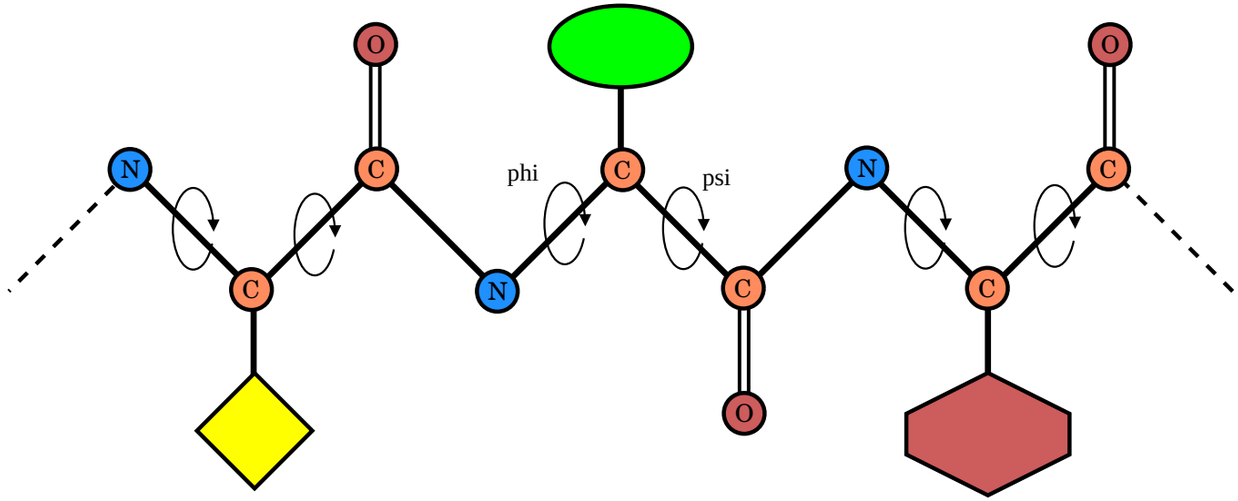
- Obvious similarity ;-)



/==



- Our model



  - Protein == vector of phi and psi angles (treelike robot with 2n dof)
  - possible models range from only backbone up to full atom

Differences to usual PRM

- no external obstacles, but
  - self-avoidingness
  - torsion angles
- quality of paths
  - low energy intermediate states
  - kinetically prefered paths
  - highly probable paths

Energy Function

- method can use any potential

Energy Function

- method can use any potential
- Our coarse potential

  [Levitt. J.Mol.Biol., 1983.]

  - each sidechain by only one “atom” (zero dof)

$$U_{tot} =$$

Energy Function

- method can use any potential
- Our coarse potential

  [Levitt. J.Mol.Biol., 1983.]

  - each sidechain by only one “atom” (zero dof)

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

  - first term favors known secondary structure through main chain hydrogen bonds and disulphide bonds

Energy Function

- method can use any potential
- Our coarse potential

  [Levitt. J.Mol.Biol., 1983.]

  - each sidechain by only one “atom” (zero dof)

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

  - first term favors known secondary structure through main chain hydrogen bonds and disulphide bonds
  - second term hydrophobic effect

Energy Function

- method can use any potential
- Our coarse potential

  [Levitt. J.Mol.Biol., 1983.]

  - each sidechain by only one “atom” (zero dof)

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

  - first term favors known secondary structure through main chain hydrogen bonds and disulphide bonds
  - second term hydrophobic effect
  - Van der Waals interaction modeled by step function

Energy Function

- method can use any potential
- Our coarse potential

  [Levitt. J.Mol.Biol., 1983.]

  - each sidechain by only one “atom” (zero dof)

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

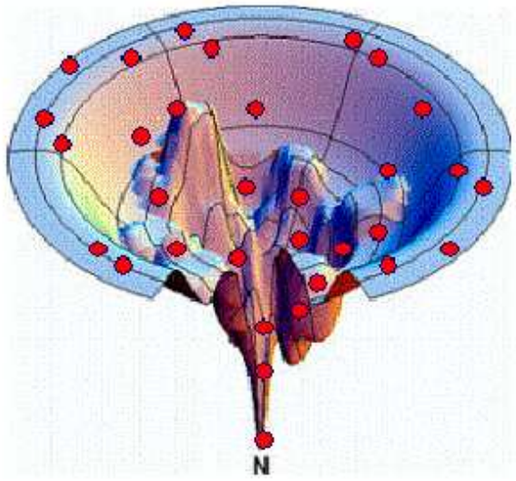
  - first term favors known secondary structure through main chain hydrogen bonds and disulphide bonds
  - second term hydrophobic effect
  - Van der Waals interaction modeled by step function
- All-atom potential: EEF1

  [Lazaridis, Karplus. Proteins, 1999.]

PRM method for Proteins

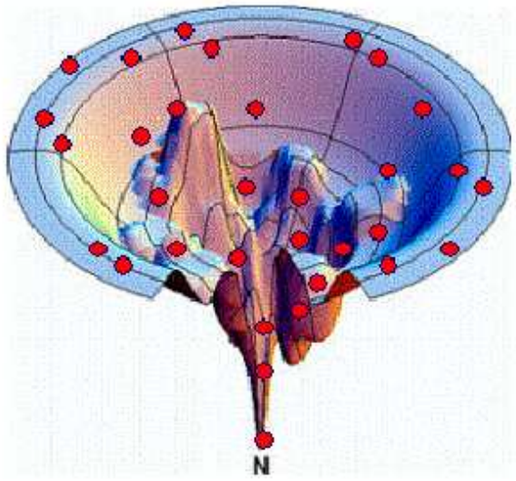
1. Motivation
2. Motion Planning
3. **Protein Model**
4. Roadmaps
5. Results
6. Conclusion

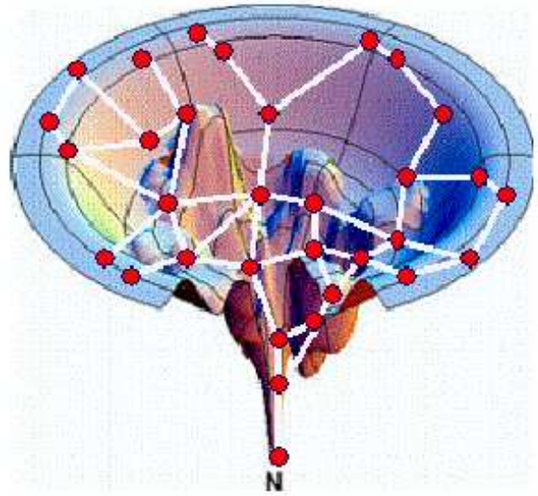
PRM method for Proteins



- Sampling

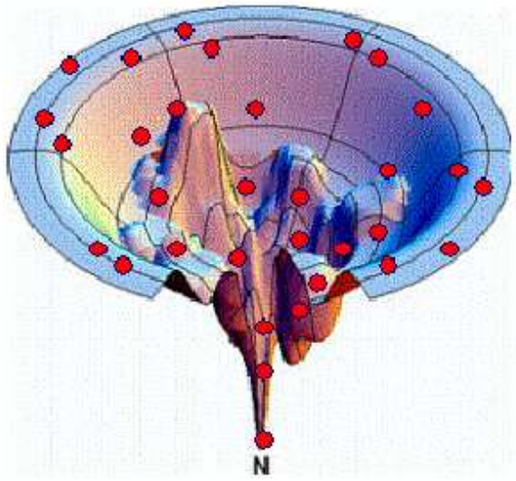
PRM method for Proteins



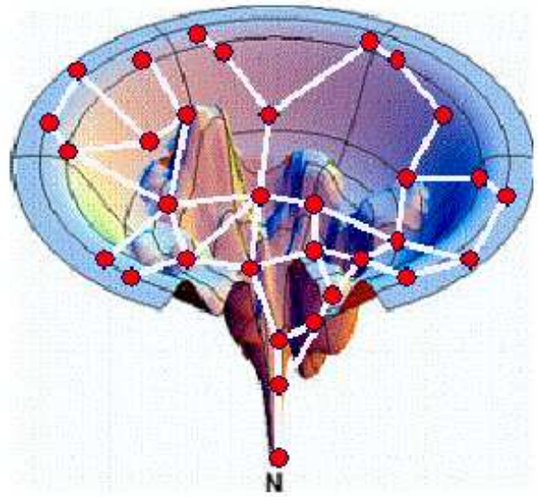


- Sampling
- Connecting

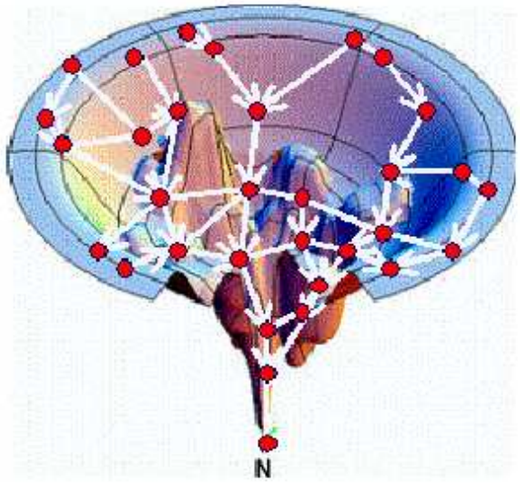
PRM method for Proteins



- Sampling

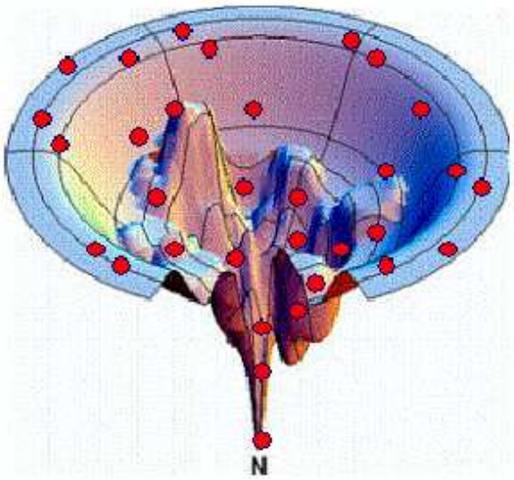


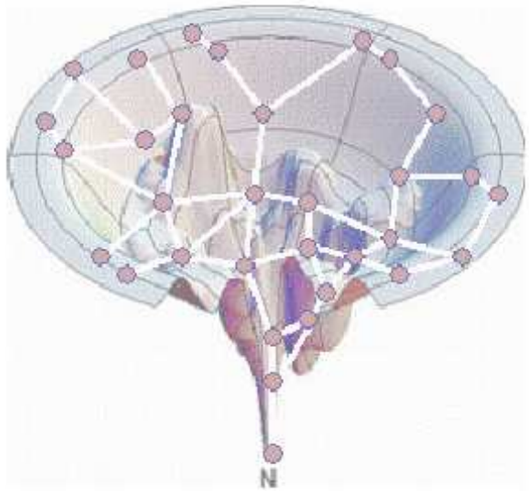
- Connecting

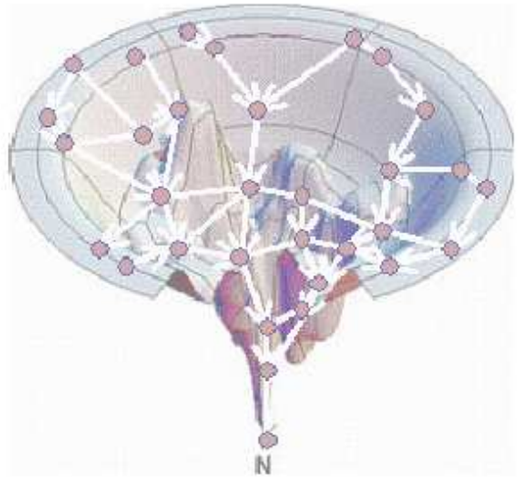


- Extracting

Sampling — Node Generation







- Sampling
- Connecting
- Extracting

Node Generation

- No uniform sampling
  - configuration space too large
  - $\Rightarrow$ need biased sampling strategy

Node Generation

- No uniform sampling
  - configuration space too large
  - $\Rightarrow$ need biased sampling strategy
- Gaussian sampling
  - centered around native conformation
  - with different STDs $5°, 10°, \ldots, 160°$
  - ensure representants for different numbers of native contacts

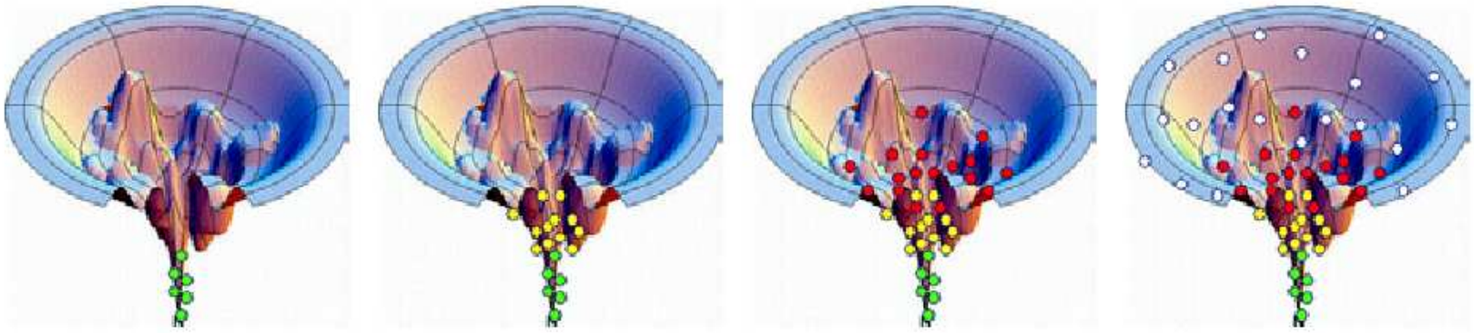
Node Generation

- No uniform sampling
  - configuration space too large
  - $\Rightarrow$ need biased sampling strategy
- Gaussian sampling
  - centered around native conformation
  - with different STDs $5°, 10°, \ldots, 160°$
  - ensure representants for different numbers of native contacts
- Selection by energy

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

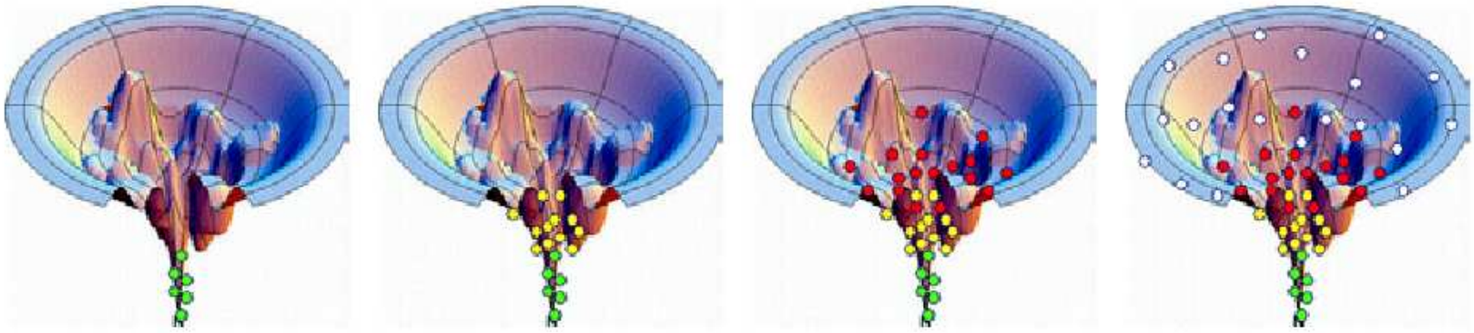
More on Node Generation

- Visualization of Sampling Strategy

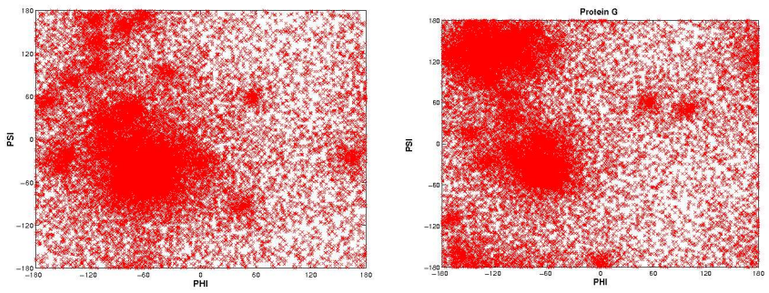


More on Node Generation

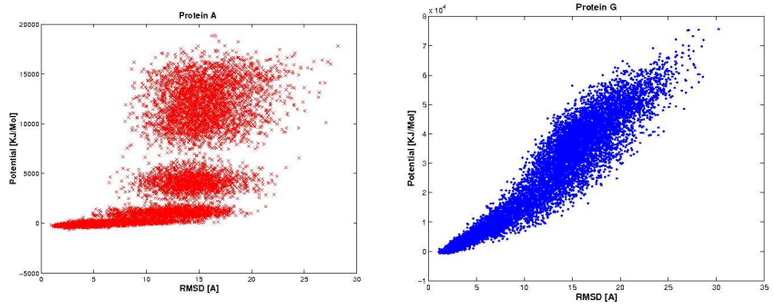
- Visualization of Sampling Strategy



- Distribution

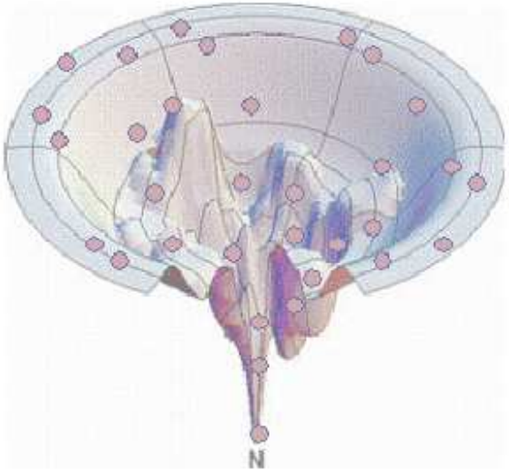


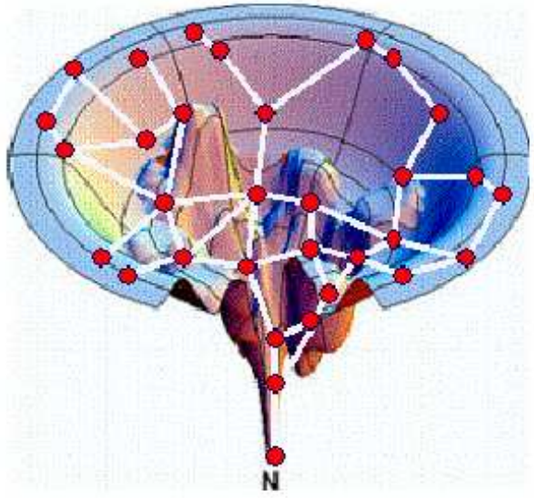
Psi and Phi angles

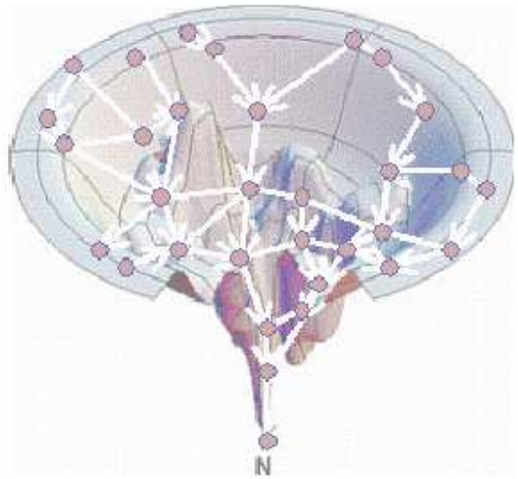


RMSD vs. Energy

Node Connection







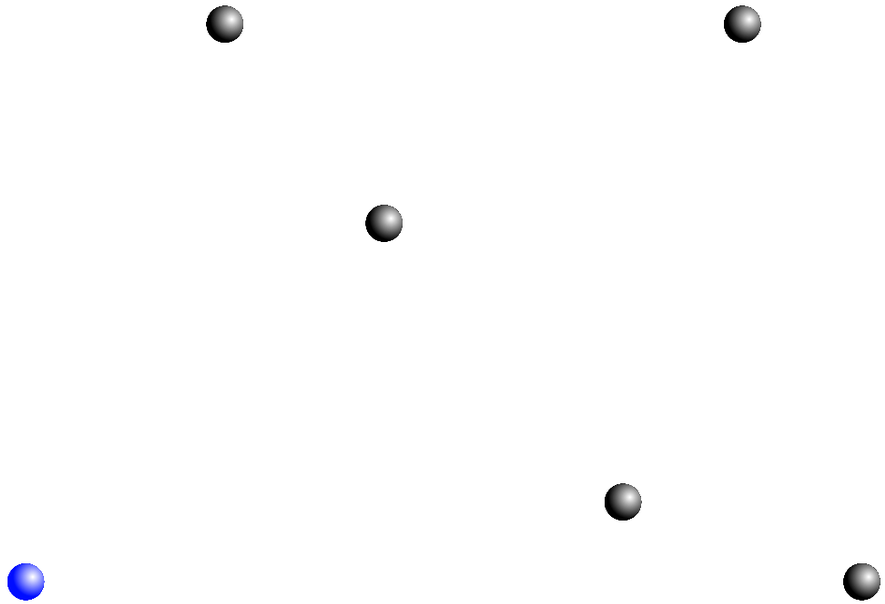
- Sampling
- Connecting
- Extracting

Connecting Nodes by Local Planner

- connect confi gurations in close distance
- generate N intermediary nodes by local planner

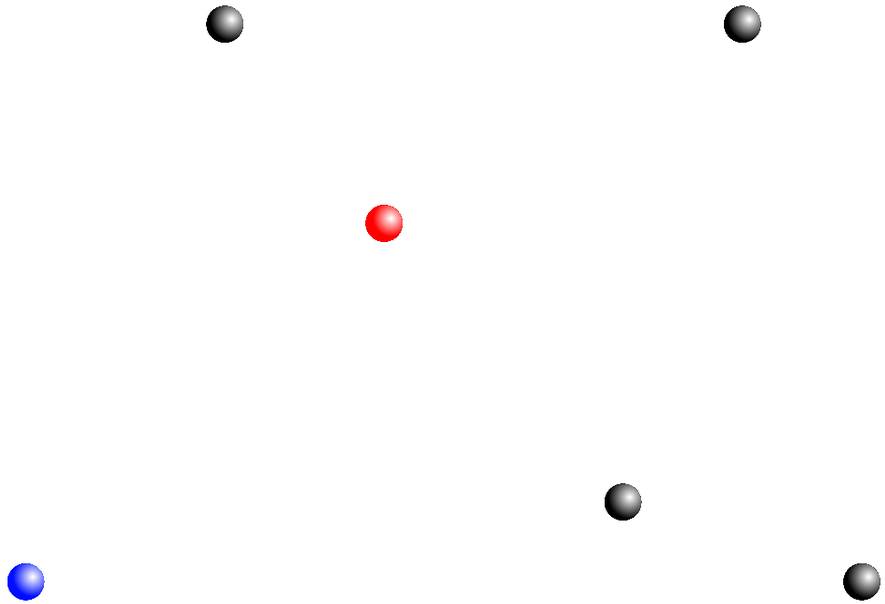
Connecting Nodes by Local Planner

- connect confi gurations in close distance
- generate N intermediary nodes by local planner



Connecting Nodes by Local Planner

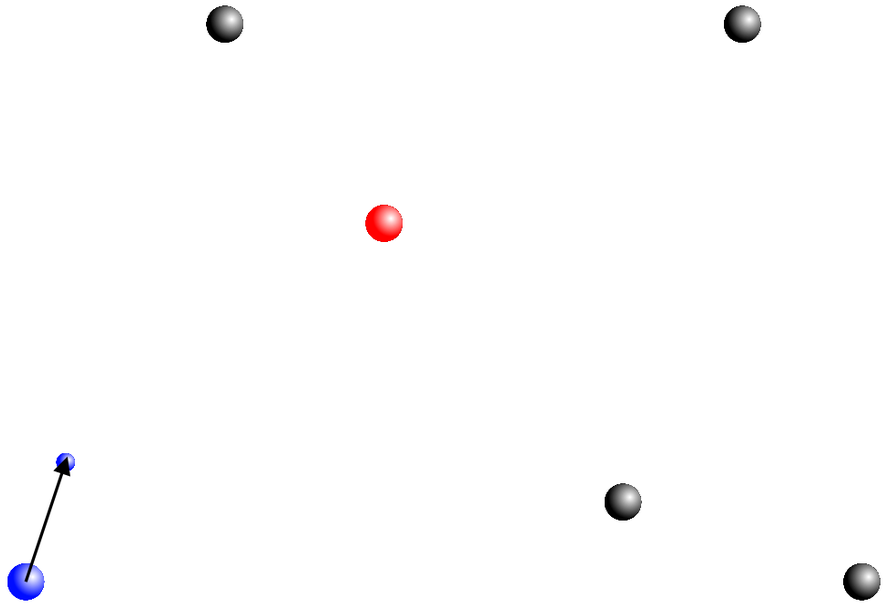
- connect confi gurations in close distance
- generate N intermediary nodes by local planner



Connecting Nodes by Local Planner

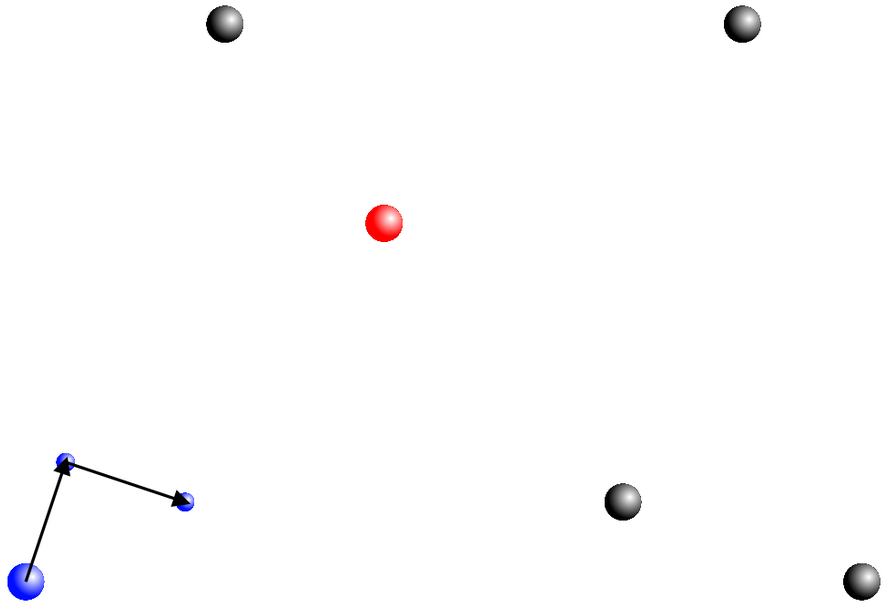
1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

- connect confi gurations in close distance
- generate N intermediary nodes by local planner



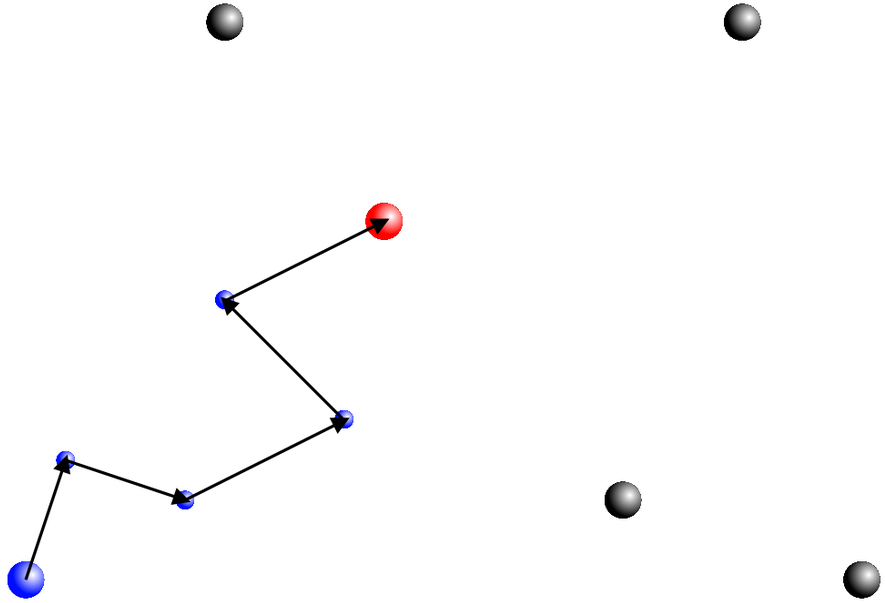
Connecting Nodes by Local Planner

- connect confi gurations in close distance
- generate N intermediary nodes by local planner



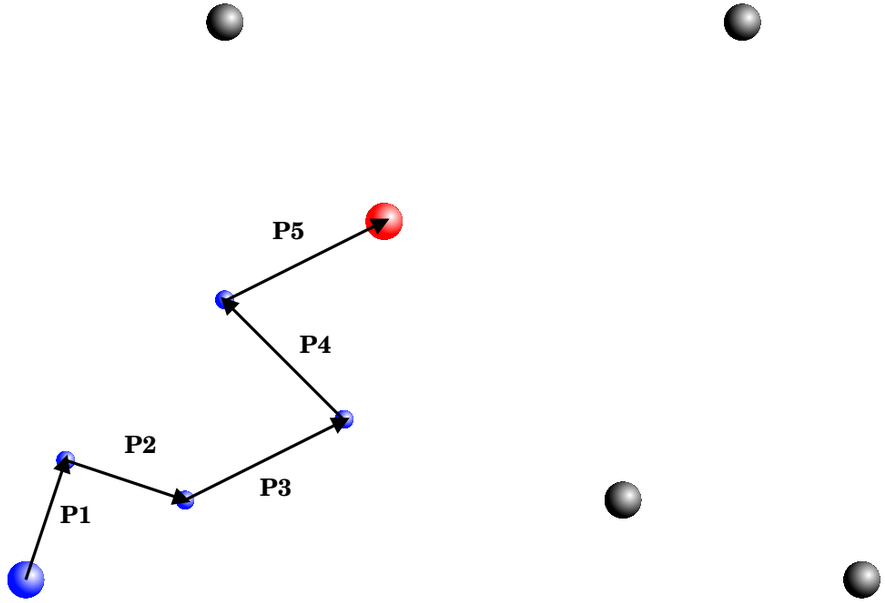
Connecting Nodes by Local Planner

- connect confi gurations in close distance
- generate N intermediary nodes by local planner



Connecting Nodes by Local Planner

- connect confi gurations in close distance
- generate N intermediary nodes by local planner

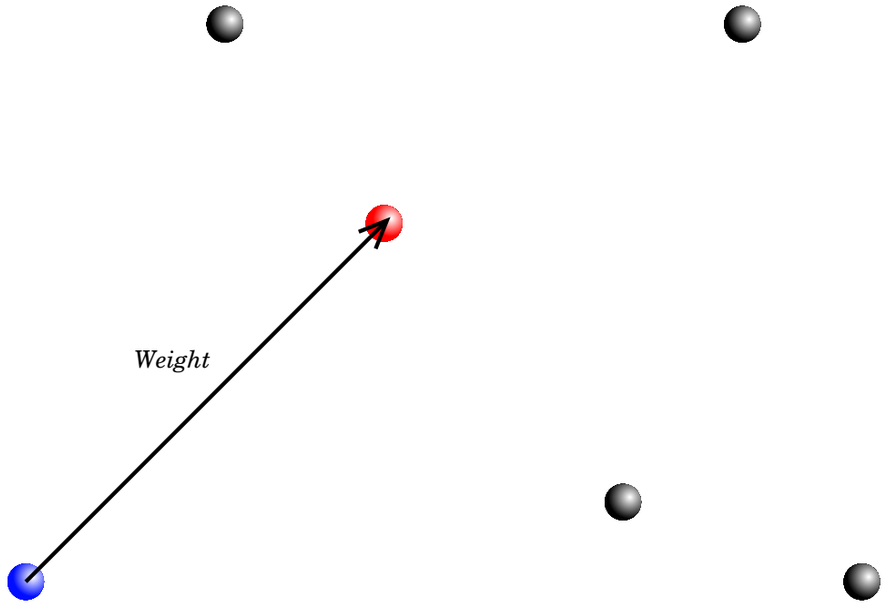


- assign weights to edges

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

Connecting Nodes by Local Planner

- connect confi gurations in close distance
- generate N intermediary nodes by local planner



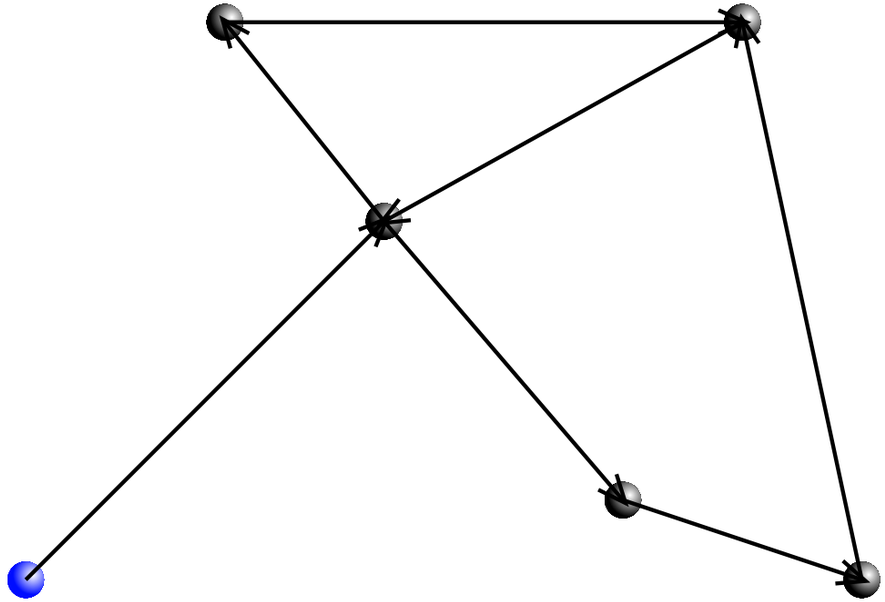
- assign weights to edges

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

$$Weight = \sum_{i=0}^{N} -log(P_i)$$

Connecting Nodes by Local Planner

- connect confi gurations in close distance
- generate N intermediary nodes by local planner

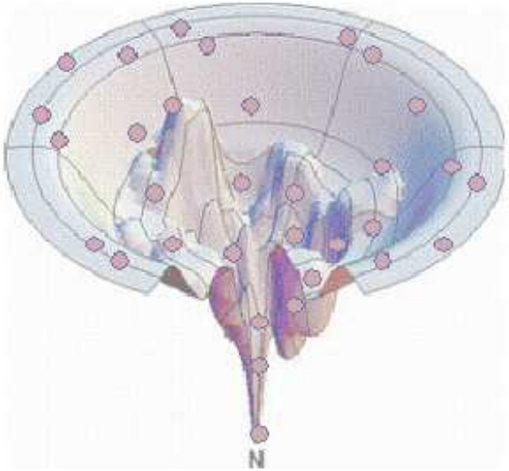


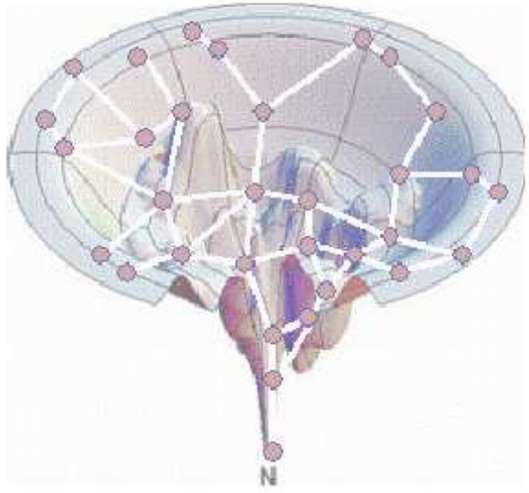
- assign weights to edges

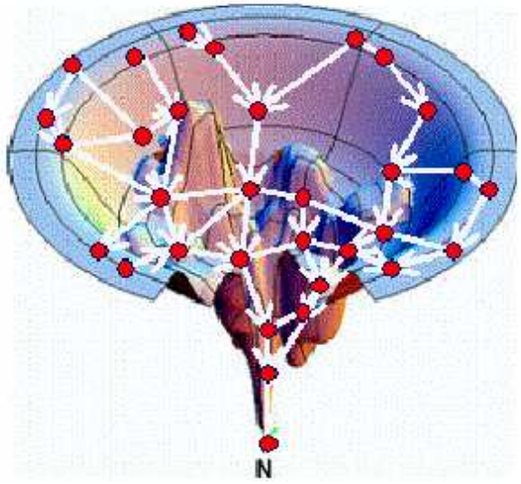
$$P_i = \begin{cases} e^{-\frac{\Delta E}{kT}} & \text{if } \Delta E > 0 \\ 1 & \text{if } \Delta E \leq 0 \end{cases}$$

$$Weight = \sum_{i=0}^{N} -log(P_i)$$

Extracting Paths







- Sampling
- Connecting
- Extracting

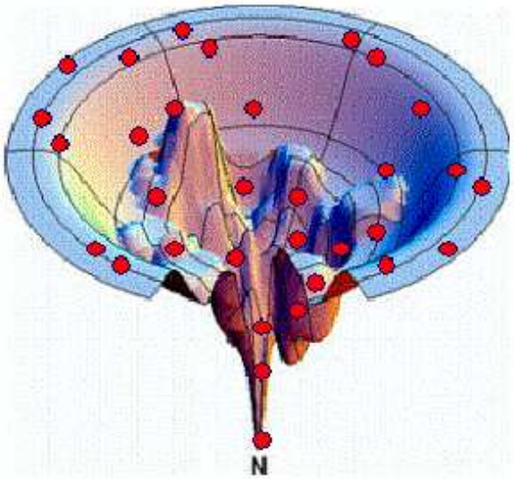
Extracting Paths

- Shortest Path
  - extract one shortest path
  - from some starting conformation, one path at a time

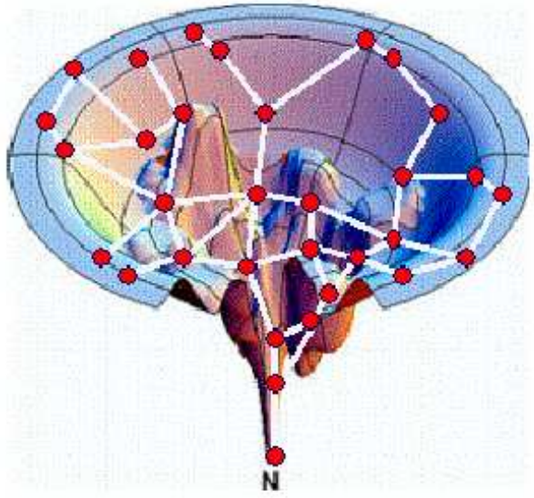
Extracting Paths

- Shortest Path
  - extract one shortest path
  - from some starting conformation, one path at a time
- Single Source Shortest Paths (SSSP)
  - extract shortest paths from all starting conformation
  - compute paths simultaneously
  - generate tree of shortest paths (SSSP tree)

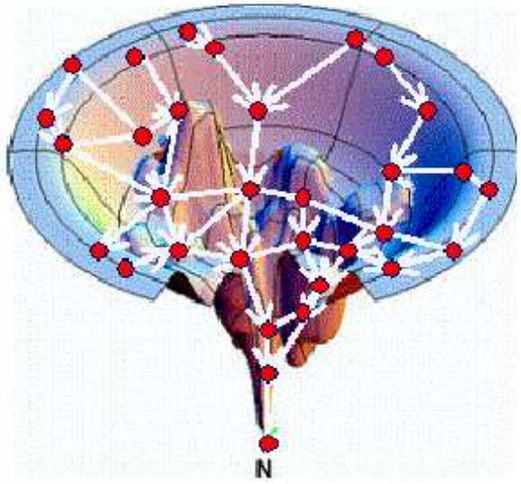
Big Picture



- Sampling



- Connecting



- Extracting

Studied Proteins

- Overview of studied proteins, roadmap size, and construction times

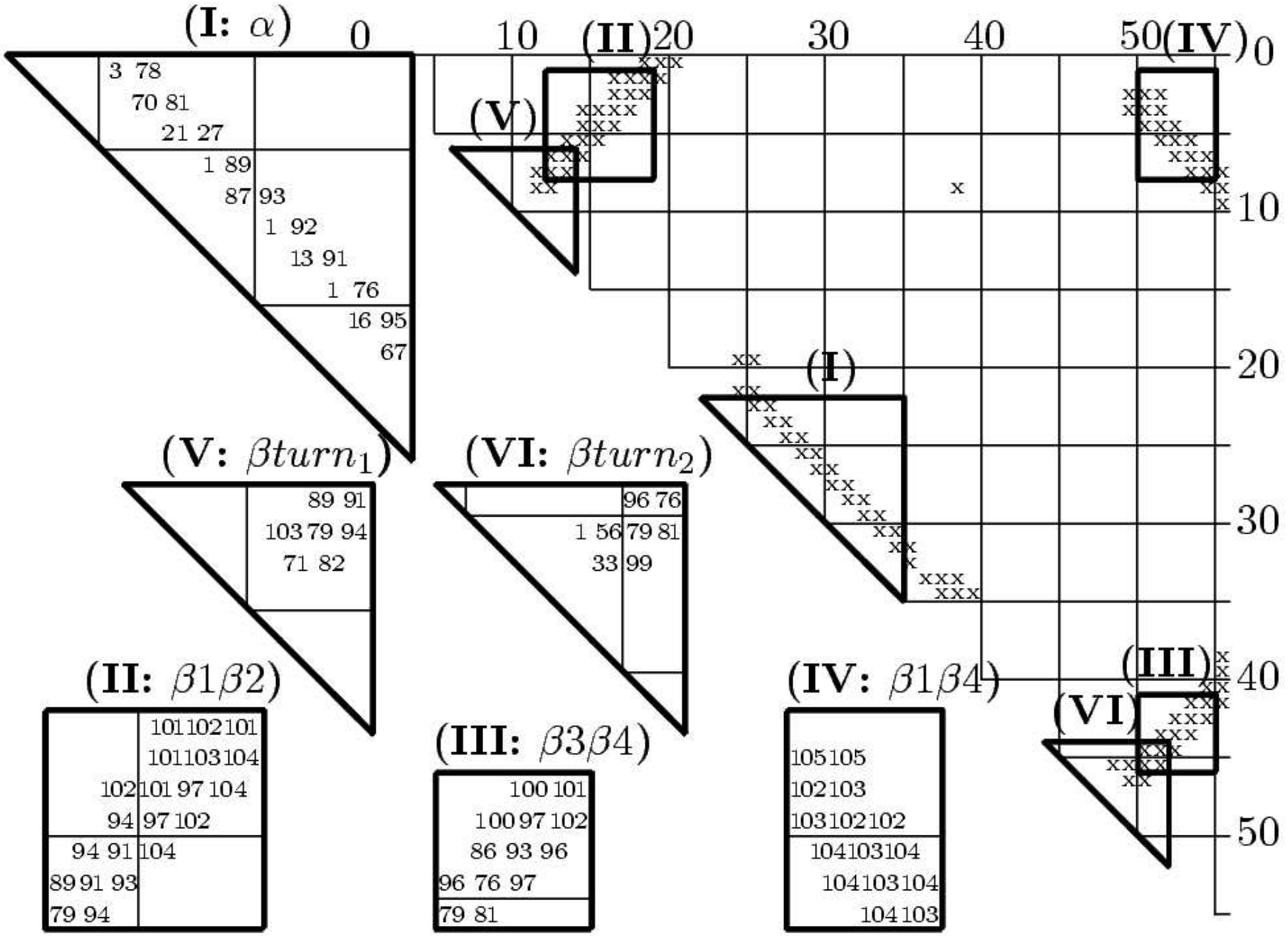
pdb	Description	Length	SS	# Nodes	Time (h)
1gb1	Protein G domain B1	56	$1\alpha + 4\beta$	8 000	6.400
2crt	Cardiotoxin III	60	5β	8 000	6.430
1bdd	Staphylococcus protein A	60	3α	10 000	10.400
1shg	SH3 domain α-spectrin	62	5β	10 000	8.344
2ptl	Protein L, B1 domain	62	$1\alpha + 4\beta$	4 000	3.104
1coa	CI2	64	$1\alpha + 4\beta$	10 000	9.984
1srl	SH3 domain src	64	5β	8 000	5.990
1nyf	SH3 domain fyn	67	5β	10 000	8.418
2ait	Tendamistat	74	7β	10 000	13.327
1ubq	Ubiquitin	76	$1\alpha + 5\beta$	8 000	10.381
1pks	SH3 domain PI3 kinase	79	$1\alpha + 5\beta$	10 000	14.446
1pba	Procarboxypeptidase A2	81	$3\alpha + 3\beta$	8 000	10.845

Formation orders

- formation order of secondary structure for verifying method
- formation orders can be determined experimentally

  [Li, Woodward. Protein Science, 1999.]
  - Pulse labeling
  - Out-exchange
- prediction of formation orders
  - single paths
  - averaging over multiple paths (SSSP-tree)

Timed Contact Maps



Formation Order

pdb	Out exchange	Pulse labeling	Our SS formation order	Comp.
1gb1	[α, β1, β3, β4], β2	[α, β4], [β1, β2, β3]	α, β3–β4, β1–β2, β1–β4	Agreed
2crt	[β3, β4, β5], [β1, β2]	β5, β3, β4, [β1, β2]	β1–β2, β3–β4, β3–β5	Not sure
1bdd	[α2, α3], α1	[α1, α2, α3]	[α2, α3], α1, α2–α3, α1–α3	Agreed
1shg	N/A	N/A	β3–β4, β2–β3, β1–β5, β1–β2	N/A
2ptl	[α, β1, β2, β4], β3	[α, β1], [β2, β3, β4]	α, β1–β2, β3–β4, β1–β4	Agreed
1coa	[α, β2, β3], [β1, β4]	N/A	α, β3–β4, β2–β3, β1–β4	Agreed
1srl	N/A	N/A	β3–β4, β2–β3, β1–β5, β1–β2	N/A
1nyf	N/A	N/A	β3–β4, β2–β3, β1–β2, β1–β5	N/A
2ait	[β1, β2], [β3, β4, β5, β6, β7]	N/A	β1–β2, β3–β4, [β2–β5, β3–β6], β3–β5	Agreed
1ubq	[α, β1, β2], [β3, β5], β4	N/A	α, β3–β4, β1–β2, β3–β5, β1–β5	Agreed
1pks	N/A	N/A	β3–β4, β1–β5, [β1–β2, β2–β3]	N/A
1pba	N/A	N/A	[α1, α3], [β1–β2, β1–β3]	N/A

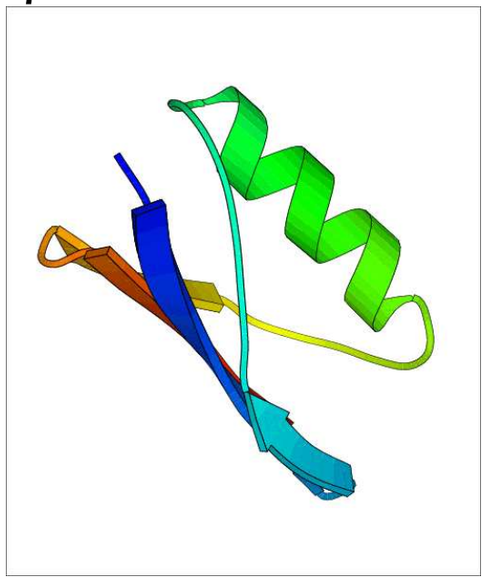
Formation Order

pdb	Out exchange	Pulse labeling	Our SS formation order	Comp.
1gb1	[α, $\beta1$, $\beta3$, $\beta4$], $\beta2$	[α, $\beta4$], [$\beta1$, $\beta2$, $\beta3$]	α, $\beta3$–$\beta4$, $\beta1$–$\beta2$, $\beta1$–$\beta4$	Agreed
2crt	[$\beta3$, $\beta4$, $\beta5$], [$\beta1$, $\beta2$]	$\beta5$, $\beta3$, $\beta4$, [$\beta1$, $\beta2$]	$\beta1$–$\beta2$, $\beta3$–$\beta4$, $\beta3$–$\beta5$	Not sure
1bdd	[$\alpha2$, $\alpha3$], $\alpha1$	[$\alpha1$, $\alpha2$, $\alpha3$]	[$\alpha2$, $\alpha3$], $\alpha1$, $\alpha2$–$\alpha3$, $\alpha1$–$\alpha3$	Agreed
1shg	N/A	N/A	$\beta3$–$\beta4$, $\beta2$–$\beta3$, $\beta1$–$\beta5$, $\beta1$–$\beta2$	N/A
2ptl	[α, $\beta1$, $\beta2$, $\beta4$], $\beta3$	[α, $\beta1$], [$\beta2$, $\beta3$, $\beta4$]	α, $\beta1$–$\beta2$, $\beta3$–$\beta4$, $\beta1$–$\beta4$	Agreed
1coa	[α, $\beta2$, $\beta3$], [$\beta1$, $\beta4$]	N/A	α, $\beta3$–$\beta4$, $\beta2$–$\beta3$, $\beta1$–$\beta4$	Agreed
1srl	N/A	N/A	$\beta3$–$\beta4$, $\beta2$–$\beta3$, $\beta1$–$\beta5$, $\beta1$–$\beta2$	N/A
1nyf	N/A	N/A	$\beta3$–$\beta4$, $\beta2$–$\beta3$, $\beta1$–$\beta2$, $\beta1$–$\beta5$	N/A
2ait	[$\beta1$, $\beta2$], [$\beta3$, $\beta4$, $\beta5$, $\beta6$, $\beta7$]	N/A	$\beta1$–$\beta2$, $\beta3$–$\beta4$, [$\beta2$–$\beta5$, $\beta3$–$\beta6$], $\beta3$–$\beta5$	Agreed
1ubq	[α, $\beta1$, $\beta2$], [$\beta3$, $\beta5$], $\beta4$	N/A	α, $\beta3$–$\beta4$, $\beta1$–$\beta2$, $\beta3$–$\beta5$, $\beta1$–$\beta5$	Agreed
1pks	N/A	N/A	$\beta3$–$\beta4$, $\beta1$–$\beta5$, [$\beta1$–$\beta2$, $\beta2$–$\beta3$]	N/A
1pba	N/A	N/A	[$\alpha1$, $\alpha3$], [$\beta1$–$\beta2$, $\beta1$–$\beta3$]	N/A

- no (reported) contradictions between prediction and validation
- different kind of information from experiment and prediction

The Proteins G and L

- Studied in more detail
- good test case
- structurally similar: $1\alpha + 4\beta$



- fold differently
  - Protein G: β-turn 2 forms first
  - Protein L: β-turn 1 forms first

Comparison of Analysis Techniques β-Turn Formation

1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

Name	Contacts considered	Energy function	Secondary structure formation order	Analyze first x% contacts				
				20	40	60	80	100
Protein G	All	Our	α, turn 2, turn 1	53	52	52	50	50
			turn 2, α, turn 1	15	9	17	22	22
			α, turn 1, turn 2	25	33	26	23	24
		All-atom	α, turn 2, turn 1	36	37	55	55	57
			turn 2, α, turn 1	3	0	0	0	0
			α, turn 1, turn 2	50	63	45	45	43
			turn 1, α, turn 2	12	0	0	0	0
	Hydrophobic	Our	α, turn 2, turn 1	96	96	85	96	87
			α, turn 1, turn 2	4	4	12	2	11
		All-atom	α, turn 2, turn 1	76	78	78	92	69
			α, turn 1, turn 2	24	22	22	8	31
Protein L	All	Our	α, turn 1, turn 2	24	30	37	38	41
			turn 1, α, turn 2	3	4	4	4	6
			α, turn 2, turn 1	73	63	60	48	39
		All-atom	α, turn 1, turn 2	25	25	48	43	41
			α, turn 2, turn 1	75	75	52	57	59
	Hydrophobic	Our	α, turn 1, turn 2	72	68	72	70	69
			turn 1, α, turn 2	5	9	5	7	15
			α, turn 2, turn 1	23	22	22	23	15
		All-atom	α, turn 1, turn 2	66	76	78	95	97
			turn 1, α, turn 2	3	0	0	0	0
			α, turn 2, turn 1	31	24	22	5	3

Conclusion

1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

Conclusion

- PRM can be applied to “realistic” protein models

Conclusion

- PRM can be applied to “realistic” protein models
- Introduced method makes verifi able prediction

Conclusion

1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

- PRM can be applied to “realistic” protein models
- Introduced method makes verifi able prediction
- Coarse potential is suffi cient

Conclusion

1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

- PRM can be applied to “realistic” protein models
- Introduced method makes verifi able prediction
- Coarse potential is suffi cient
- Predictions are in good accordance to experimental data

Conclusion

1. Motivation
2. Motion Planning
3. Protein Model
4. Roadmaps
5. Results
6. Conclusion

- PRM can be applied to “realistic” protein models
- Introduced method makes verifi able prediction
- Coarse potential is suffi cient
- Predictions are in good accordance to experimental data
- Interesting relations to e.g. computation of barrier trees