Robotics-inspired Methods for Biological Applications: From Robots to Proteins

Instructor: Amarda Shehu
amarda@gmu.edu

Today’s lecture:
Robotics-inspired Methods
Focus: Protein-central Applications
Analogies Employed by Robotics Researchers

- Protein molecule
  - spheres - atoms
  - cylinders – bonds connecting spheres

- Protein conformation
  - spatial arrangement of atoms

- Transition between conformations
  - rotation of bonds

- A protein chain connects atoms with bonds

Analogy between proteins and kinematic chains with revolute joints:

- Bonds can be treated as links
- Atoms can be treated as joints
- Rotation of bonds can be modeled as rotation of joints
- Rotations give rise to protein *conformations*
- Conformation is the equivalent of a configuration
Problems of Interest to Robotics Researchers

- A protein transitions between conformations until it finds a lowest-energy conformation where it is biologically active.
- Such a conformation is guaranteed to exist from the theory of equilibrium thermodynamics.
- The native conformation is also referred to as the folded conformation.

Interesting problems on proteins:

- Structure Prediction: Given the sequence of the atoms that make up a protein molecule, determine the native conformation.
- Folding: Given the folded conformation, find the paths that the protein follows to fold onto this conformation.
- Docking: Given a folded protein conformation and a small drug molecule, find how and where the drug docks onto the molecule.
Molecular Motion is Essential for Life

Mad cow disease is caused by misfolding

Drug molecules act by binding to proteins
Sometimes protein is too flexible and does not bind to drug
Configuration Space and Conformational Space

- **Protein conformational space**
  - articulated robot configuration space of higher dimensionality
- **Similar sampling techniques**
  - dofs - torsional angles
- **0/1 collisions**
  - energy surface instead of atoms interact with one another giving rise to potential energy

Hundreds to thousands of revolute dofs in real protein chains
Robotics-inspired Methods to Study Proteins

- Problem often explored in robotics community: Given two protein conformations A and B, plan motions from A to B
  - When B is the experimentally-available native/folded structure, sequence of computed motions offers likely “folding” pathways

- Sampling-based motion planning methods for proteins
  - Cortes J. et al. Bioinformatics 21, 2005
  - Georgiev I. and Donald, R. B. Bioinformatics 23, 2007

- Folding problem is often targeted due to its analogies with finding paths to connect a start and goal configuration in a roadmap

- Structure prediction problem, where the native conformation is not known and needs to be computed, is considered more challenging
Focus first on a Structure Prediction Approach
1. Exploration of a High-dimensional Space

- Backbone of N amino acids → 2N dofs
- Energy surface associated with conformational space
- Evolution has “guided” native state (in naturally-occurring proteins) to be lowest free-energy state
- Energy surface is rugged and constellated with local minima

In structure prediction: Explore a high-dimensional space in search of native-like (low-energy) conformations
2. Representation (Modeling Problem)

- Discrete: atoms in a lattice
- Continuous: off-lattice models

Coarse-grained to fine-grained

Needed: diverse coarse-grained conformations near the native state that can be further refined in all-atom detail

State-of-the-art approach: compute at coarse-grained detail, refine at higher resolution later
3. Vast Space, Approximate Energy Function

Empirical force-fields to measure potential energy: *AMBER ff*, CHARMM, OPLS, AMW, Rosetta, ...

- **bond stretch**
  \[ \sum_{\text{bonds}} K_b(b - b_0)^2 \]

- **torsional**
  \[ \sum_{\text{dihedrals}} K_\theta(1 + \cos(n\theta - \delta)) \]

- **valence angle bend**
  \[ \sum_{\text{angle}} K_\alpha(\alpha - \alpha_0)^2 \]

- **nonbonded**
  \[ \sum_{\text{nonbonded}} \epsilon[(R_{\text{min}}/r_{ij})^{12} - (R_{\text{min}}/r_{ij})^6] \]

Inherent errors or biases in energy functions warrant focus on diverse emerging minima relevant for the native state.
Exploration of Conformational Space

- Traditional Approaches - Monte Carlo (MC)
  - trajectory-based exploration
  - initial conformations, length of trajectories, number of trajectories are important decisions

- Strategies to enhance MC sampling:

- State-of-the-art on enhancing sampling of native-like conformations:
  - fragment-based assembly on simpler coarse-grained conformational space

Challenge: ensuring computed conformations are geometrically-distinct and not representative of only a few regions of conformational space

- Proposed method: **Fragment Monte Carlo Tree Exploration (FeLTr)**

- **Goal**: rapidly compute diverse native-like conformations
  - applicability: conformations can serve as good starting points for larger detailed studies of protein engineering and design

- **Novelty**: tree-based exploration guided with projection layers
Tree-based Exploration in FeLTr

- Tree-based search
- (i) select vertex for expansion
- (ii) expand vertex

(i) reconciles
- towards lower energies
- towards diverse conformations

(ii) – Metropolis Monte Carlo trajectory
- employs fragment-based assembly
- move = fragment configuration found in native conformations

Tree naturally integrates decisions about number of trajectories and selection of conformations in a trajectory from where to continue the exploration.
Expansion Step in FeLTr: Monte Carlo Trajectory

Assemble conformations with physical configurations of protein fragments

- Fragment = k consecutive amino acids
- Configurations of this fragment: 6 backbone dihedral angles
- Sample configurations from those found in native protein structures for a fragment of that specific amino-acid sequence

Monte Carlo move in FeLTr: propose configuration of k = 3 three consecutive amino-acids in conf

Metropolis criterion: change in energy evaluated after an attempted move, accepted with probability \(e^{-\Delta E/RT}\)

Expanding C through N-2 moves results in Cnew

Database of trimer configurations built from non-redundant PDB

\(20^3\)
Expansion Step: Representation and Energy Function

- CB representation: explicit backbone and CB atoms
- Energy evaluated with coarse energy function based on AMW of Papoian et al. PNAS 101, 2004
- Additional energy term to obtain compact conformations
- Purpose: native-like conformations are compact
- Penalizes conformation if radius of gyration $R_g$ (root-mean-square distance of atoms from c.o.m.) > $R^*$ value
- $R^*$ value calculated from what is expected for a chain of same number of amino acids in the PDB
  - $R_{g_{PDB}} = 2.83 \times N^{0.34}$
Selection Step in FeLTr: Two Projection Layers

Select for expansion low-energy conformations that fall in under-explored regions of the conformational space

- Tree mapped on:
  - (i) 1d energy grid
  - (ii) 3d grid – projection of conformational space

- Grids used to bias selection for expansion to:
  - conformations that have low energies and
  - map to scarce cells of projection space
Geometric Projection of Conformational Space

Finding a few conformational coordinates (reaction coordinates) to effectively represent and compare conformations is an open research area.

- Ultrafast shape recognition (USR) features proposed to encapsulate overall shape [Ballester, Richards, J. Comput. Chem. 2007]
- Features are momenta of atomic distance distributions from four reference points:
  - centroid (ctd)
  - closest to centroid (cst)
  - farthest from cst (fct)
  - farthest from fct (ftf)
- Reference points capture well-separated extremes in a conformation
- Atomic distances from each reference point yield non-redundant information
To select vertex for expansion
1. select energy level \( l \)
2. select projection cell \( c \)

1. Select energy level \( l \)
   \[ w(l) = E_{\text{avg}}(l) * E_{\text{avg}}(l) + \varepsilon \]
   \[ p(l) = w(l) / \left[ \Sigma_{l'} w(l') \right] \]

2. Select projection cell \( c \)
   \[ w(c) = 1.0 / \left[ (1 + nsel(c)) * nconfs(c) \right] \]
   \[ p(c) = w(c) / \left[ \Sigma_{c'} w(c') \right] \]

Details:
\( \varepsilon = 1.0/2^{22} \)
\( nsel(c) = \text{nr. times } c \text{ selected} \)
\( nconfs(c) = \text{nr. confs in } c \)
Chosen Test Proteins and Simulation Settings

<table>
<thead>
<tr>
<th>Protein</th>
<th>Trp-cage</th>
<th>wWD</th>
<th>hp36</th>
<th>eHD</th>
<th>L20</th>
<th>GB1</th>
<th>Calbindin D9k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold</td>
<td>α</td>
<td>β</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>α/β</td>
<td>α</td>
</tr>
<tr>
<td>Size</td>
<td>20</td>
<td>26</td>
<td>36</td>
<td>54</td>
<td>60</td>
<td>60</td>
<td>76</td>
</tr>
<tr>
<td>Dofs</td>
<td>40</td>
<td>52</td>
<td>72</td>
<td>108</td>
<td>120</td>
<td>120</td>
<td>152</td>
</tr>
</tbody>
</table>

Tryptophan cage, Pin1 Trp-Trp ww domain, villin headpiece, engrailed homeo-domain, bacterial ribosomal protein, immunoglobulin binding domain of streptococcal protein G, calbindin D9k

- **Implementation & system details:**
  - 3d grid: 30x30x30
  - 1d grid cell size: 2 kcal/mol
  - C++, Intel Core2 Duo, 4GB RAM, 2.66 GHZ CPU

- **Simulation settings:**
  - capture native state in limited simulation time: <= 3 CPU hours
  - fit exploration tree and ensemble in memory: <= 50,000 conformations

- **Analysis of applications on the seven chosen test sequences focuses on:**
  - is experimentally-available native structure reproduced?
  - what is the diversity of ensemble of lowest-energy conformations?
  - compare with a Monte Carlo simulation of same *fragment assembly*, *energy function*, *time* and *ensemble size*
Able to Capture Diverse Minima in Limited Time

- Analysis shows better use of time and population of more energy minima than an MC simulation
- Some native topologies difficult to capture with current energy function
Focus on Folding

Given the folded conformation, compute folding pathways that the protein potentially follows to fold onto the biologically-active/native conformation soon after it is synthesized in the ribosome
Roadmap-Based Representation

- Compact representation of many motion pathways
- Coarse resolution relative to MC simulations
- Efficient algorithms for analyzing multiple pathways
Roadmaps on Protein Folding

- Known native conformation of a protein chain
- Degrees of freedom: $\varphi$-$\psi$ angles
- Energy: van der Waals, hydrogen bonds, hydrophobic effect
- **New idea**: Novel sampling to generate conformations
- **Application**: Finding order of formation of certain secondary structures that can be validated with experimental data
Sampling Strategy (Node Generation)

- High dimensionality
  → non-uniform sampling

- Conformations generated by sampling angle values from Gaussian distributions around angles of reference conformations

- Two-tier sampling employed to obtain good coverage of conformational space fast
Sampling Strategy (Node Generation)

- **Tier 1**: Gaussian distribution around native state

- Conformations are sorted into bins by number of native contacts (pairs of neighboring Cα atoms in native conformation)

- **Tier 2**: Conformations from low-filled bins picked to be perturbed
  - Gaussian distributions with increasing standard deviations
Sampling Strategy (Node Generation)

- A generated conformation is added to the roadmap with probability \( P \):
  \[
  P = \begin{cases} 
  0 & \text{if } E > E_{\text{max}} \\
  \frac{E_{\text{max}} - E}{E_{\text{max}} - E_{\text{min}}} & \text{if } E_{\text{min}} \leq E \leq E_{\text{max}} \\
  1 & \text{if } E < E_{\text{min}} 
  \end{cases}
  \]

- Values of \( E_{\text{min}} \) and \( E_{\text{max}} \) chosen to obtain conformations with few collisions (well-separated side-chain spheres)
Simple straight-line planner is employed to connect two neighboring conformations A and B.

Euclidean distance used to measure distance(A, B).

A fixed number of intermediate conformations in a straight line checked for energetic feasibility.
- All intermediates have to be feasible to connect A and B with an edge.
- Sum of negative logs used to associate a weight with the edge.
Application: Order of Formation of Secondary Structure Elements (SSE)

- The lowest-weight path is extracted from each denatured conformation to the folded one.
- The order of formation of SSE’s is computed along each path.
- The formation order that appears the most often over all paths is considered the SSE formation order of the protein.
A summary contact matrix is constructed, which records the time step when each native contact appears.

The time step at which a secondary structure element appears is approximated as the average of the appearance time steps of the contacts that participate in the structure.
Protein CI2
(1α + 4 β)

α forms at time step 122 (II)
β3 and β4 come together at 187 (V)
β2 and β3 come together at 210 (IV)
β1 and β4 come together at 214 (I)
α and β4 come together at 214 (III)
## Comparison with Experimental Data

<table>
<thead>
<tr>
<th>pdb</th>
<th>res. #</th>
<th>SSE’s</th>
<th>secondary structure formation order</th>
<th>roadmap size</th>
<th>exp.[25]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GB1</td>
<td>56</td>
<td>1α+4β</td>
<td>α1, β3-β4, β1-β2, β1-β4</td>
<td>5126, 70k</td>
<td>Agreed</td>
</tr>
<tr>
<td>1BDD</td>
<td>60</td>
<td>3α</td>
<td>α2, α3, α1, α2-α3, α1-α3</td>
<td>5471, 104k</td>
<td>Agreed</td>
</tr>
<tr>
<td>1COA</td>
<td>64</td>
<td>1α+4β</td>
<td>α1, β3-β4, β2-β3, β1-β4, α1-β4</td>
<td>7975, 104k</td>
<td>Agreed</td>
</tr>
<tr>
<td>1UBQ</td>
<td>76</td>
<td>1α+5β</td>
<td>α1, β3-β4, β1-β2, β3-β5, β1-β5</td>
<td>8357, 119k</td>
<td>Agreed</td>
</tr>
</tbody>
</table>

Contact formation orders from hydrogen-exchange experiments used for validation
Focus on Mini-version of Structure Prediction

Goal: analyze flexibility of specific fragments in protein chains
fragment of interest shown in blue
Why: fragment may impact interaction of protein with a drug molecule
Conformational Analysis of Protein Loops

New idea:
Explore the clash-free subset of the conformational space of a loop through an RRT-based method

Kinematic model: $\phi$-$\psi$ angles on the backbone + $\chi_i$ torsional angles in side chains
Amylosucrase (AS)
- Only enzyme in its family that acts on sucrose substrate
- The 17-residue loop (named loop 7) between Gly433 and Gly449 is believed to play a pivotal role
RRT Construction in RLG

- Extending RRT rooted at a start conformation \( q_{\text{start}} \):
  - A loop conformation \( q_{\text{rand}} \) is generated at random
    - Need satisfy neither closure nor clash-free check
  - Node \( q_{\text{near}} \) nearest to \( q_{\text{rand}} \) in current RRT is selected
    - \( q_{\text{near}} \) satisfies both closure and clash-free check
  - New nodes \( q_{\text{feas}} \) are obtained by iteratively pulling \( q_{\text{near}} \) towards \( q_{\text{rand}} \)
  - A new node \( q_{\text{new}} \) that is an intermediate between \( q_{\text{near}} \) and \( q_{\text{rand}} \) is added to the RRT
RRT Construction in RLG

- $q_{\text{rand}}$
- $q_{\text{near}}$
- $q_{\text{start}}$
RRT Construction in RLG

Stops when one can’t get closer to $q_{\text{rand}}$ or a clash is detected
Computational Results

- Surprisingly, loop 7 can’t move much
- Main bottleneck is residue Asp231
Computational Results

- Surprisingly, loop 7 can’t move much
- Main bottleneck is residue Asp231
Computational Results

- If residue Asp231 is “removed”, then loop 7’s mobility increases dramatically. The C\(\alpha\) atom of Ser441 can be displaced by more than 9Å from its crystallographic position.
Focus on Docking:
Find protein-ligand conformations with lowest interaction energy

- Study of ligand-protein binding
- Ligand is a small flexible molecule, but protein is assumed rigid
- Problem is to find configurations of the ligand near the protein’s unknown active site that result in low-energy protein-ligand pairs
- PRM-based method
  - Goal protein-ligand state is not known
  - Goal states need to be sampled as well
A fixed coordinate system P is attached to the protein.

A moving coordinate system L is attached to the ligand. L is defined using three bonded atoms in the ligand.

A conformation of the ligand is defined by the position and orientation of L relative to P and the torsional angles of the ligand.
Roadmap Construction (Node Generation)

• The nodes of the roadmap are generated by sampling conformations of the ligand uniformly at random in the parameter space (around the protein).

• The energy $E$ at each sampled conformation is computed:

\[
E = E_{\text{interaction}} + E_{\text{internal}}
\]

\[
E_{\text{interaction}} = \text{electrostatic} + \text{van der Waals potential}
\]

\[
E_{\text{internal}} = \sum_{\text{non-bonded pairs of atoms}} \text{electrostatic} + \text{van der Waals}
\]
Roadmap Construction (Node Generation)

- The nodes of the roadmap are generated by sampling conformations of the ligand uniformly at random in the parameter space (around the protein).
- The energy $E$ at each sampled conformation is computed:
  \[
  E = E_{\text{interaction}} + E_{\text{internal}}
  \]
  \[
  E_{\text{interaction}} = \text{electrostatic} + \text{van der Waals potential}
  \]
  \[
  E_{\text{internal}} = \sum_{\text{non-bonded pairs of atoms}} \text{electrostatic} + \text{van der Waals}
  \]
- A sampled conformation is retained as a node of the roadmap with probability:
  \[
  P = \begin{cases} 
  0 & \text{if } E_{\min} \leq E \leq E_{\max} \\
  \frac{E_{\max} - E}{E_{\max} - E_{\min}} & \text{if } E_{\min} \leq E \leq E_{\max} \\
  1 & \text{if } E < E_{\min}
  \end{cases}
  \]

→ Denser distribution of nodes in low-energy regions of conformational space
Roadmap Construction (Edge Generation)

- Each node connected to closest neighbors by straight edges
- Each edge is discretized so that between $q_i$ and $q_{i+1}$ no atom moves by more than some $\varepsilon$ (= 1Å)

If any $E(q_i) > E_{\text{max}}$, then the edge is rejected
Roadmap Construction (Edge Generation)

- Any two nodes closer than some threshold distance are connected by a straight edge.
- Each edge is discretized so that between \( q_i \) and \( q_{i+1} \) no atom moves by more than some \( \varepsilon \) (\( = 1\text{Å} \)).
- If for all \( q_i \), \( E(q_i) \leq E_{\text{max}} \), the edge \( q \) to \( q' \) is retained and is assigned two weights \( w(q \rightarrow q') \) and \( w(q' \rightarrow q) \):

\[
\begin{align*}
\text{Heuristic measure of energetic difficulty of moving from } q \text{ to } q' & = \\
\sum_i -\ln(P[q \rightarrow q_{i+1}])
\end{align*}
\]

where:

\[
P[q \rightarrow q_{i+1}] = \frac{e^{-(E_{i+1}-E_i)/kT}}{e^{-(E_{i+1}-E_i)/kT} + e^{-(E_{i-1}-E_i)/kT}}
\]

(probability that the ligand moves from \( q_i \) to \( q_{i+1} \) when it is constrained to move along the edge)
Querying the Roadmap

- For a given goal node $q_g$ (e.g., binding conformation), compute lowest-weight paths from $q_g$ to each node (in either direction) in $O(N \log N)$ time, where $N =$ number of nodes.

- Various quantities can then be easily computed in $O(N)$ time, e.g., average weights of all paths entering $q_g$ and of all paths leaving $q_g$ ($\sim$ binding and dissociation rates $K_{on}$ and $K_{off}$).

Protein: Lactate dehydrogenase
Ligand: Oxamate (7 degrees of freedom)
Experiments on 3 Protein-ligand Complexes

1) PDB ID: 1ldm
   Receptor: Lactate Dehydrogenase (2386 atoms, 309 residues)
   Ligand: Oxamate (6 atoms, 7 dofs)

2) PDB ID: 4ts1
   Receptor: Mutant of tyrosyl-transfer-RNA synthetase (2423 atoms, 319 residues)
   Ligand: L- leucyl-hydroxylamine (13 atoms, 9 dofs)

3) PDB ID: 1stp
   Receptor: Streptavidin (901 atoms, 121 residues)
   Ligand: Biotin (16 atoms, 11 dofs)
Computation of Potential Binding Conformations

1) Sample many (several 1000’s) ligand conformations at random around protein

2) Repeat several times:
   - Select lowest-energy conformations that are close to protein surface
   - Resample around them

1) Retain k (~10) lowest-energy conformations whose centers of mass are at least 5Å apart
Results for 1ldm

- Some potential binding sites have slightly lower energy than the active site
  → Energy is not a discriminating factor

- Average path weights (energetic difficulty) to enter and leave binding site are significantly greater for the active site
  → Indicates that active site is surrounded by energy barrier that “traps” the ligand

<table>
<thead>
<tr>
<th>Row number</th>
<th>RMSD from active site configuration (Å)</th>
<th>Configuration energy (kcal/mol)</th>
<th>Avg path weight entering configuration</th>
<th>Avg path weight leaving configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00</td>
<td>-11.79</td>
<td>112.98</td>
<td>134.54</td>
</tr>
<tr>
<td>1</td>
<td>31.04</td>
<td>-13.65</td>
<td>85.07</td>
<td>109.94</td>
</tr>
<tr>
<td>2</td>
<td>27.49</td>
<td>-12.66</td>
<td>90.48</td>
<td>111.98</td>
</tr>
<tr>
<td>3</td>
<td>1.73</td>
<td>-11.72</td>
<td>113.81</td>
<td>137.28</td>
</tr>
<tr>
<td>4</td>
<td>28.99</td>
<td>-11.54</td>
<td>85.32</td>
<td>105.19</td>
</tr>
<tr>
<td>5</td>
<td>24.67</td>
<td>-11.31</td>
<td>86.26</td>
<td>103.95</td>
</tr>
<tr>
<td>6</td>
<td>29.84</td>
<td>-11.27</td>
<td>86.49</td>
<td>107.53</td>
</tr>
<tr>
<td>7</td>
<td>29.32</td>
<td>-11.04</td>
<td>85.24</td>
<td>104.64</td>
</tr>
<tr>
<td>8</td>
<td>27.07</td>
<td>-10.96</td>
<td>81.70</td>
<td>102.28</td>
</tr>
<tr>
<td>9</td>
<td>31.00</td>
<td>-10.13</td>
<td>87.69</td>
<td>104.50</td>
</tr>
<tr>
<td>10</td>
<td>28.24</td>
<td>-9.97</td>
<td>86.36</td>
<td>98.89</td>
</tr>
</tbody>
</table>
Energy Mapping of Potential Binding Sites

Prediction where the ligand binds/docks onto the protein
Conclusion for Robotics-inspired Structure Prediction and Folding Methods on Proteins

- Probabilistic sampling (roadmaps and trees) is a recent but promising tool for exploring conformational space and computing properties of molecular pathways.

- Current/future research:
  - Better **sampling strategies** able to handle more complex molecular models (protein-protein binding)
  - Initial work on including **time information** in roadmaps
  - More thorough experimental validation to compare computed and measured **quantitative properties**