Protein Design on the Illinois Bio-Grid

Outline

- Central Dogma of Molecular Biology
- Protein Structure
- Protein Design
- Computational Challenges
- The Illinois Bio-Grid
Protein Design on the Illinois Bio-Grid

Central Dogma of Molecular Biology

- Described by Crick in 1958

- Describes how information contained in the DNA of an organism is ultimately expressed

- "The central dogma is likely to prove a considerable over-simplification"
Protein Design on the Illinois Bio-Grid
Protein Design on the Illinois Bio-Grid

Proteins

- Proteins are responsible for nearly every life process in an organism.
  - Structural components
  - Enzymes
  - Transportation
  - Signaling and intercellular communication
  - Much more
Protein Design on the Illinois Bio-Grid

Protein Structure
- Proteins are composed of a sequence of amino acids.
- The “backbone” of the protein repeats C-C-N throughout the entire protein.
- The “side chains” of the protein determine the identity of the amino acid.
Protein Design on the Illinois Bio-Grid

Amino Acids
- There are 20 amino acids
- Each amino acid has different bio-chemical properties
  - Hydrophobic / Hydrophilic
  - Charged
  - Acidic
  - ..... 

Threonin
Asparagine
Tryptophan
Protein Design on the Illinois Bio-Grid

The combined biochemical properties of the amino acids determine the three-dimensional shape of the protein.

The shape of a protein determines its effect on its environment.

http://www-vis.lbl.gov/Research/ProteinShop/
Protein Design on the Illinois Bio-Grid

What if we could redesign existing proteins with increased or additional functionality?

What if we could design proteins with novel functionality?
Protein Design on the Illinois Bio-Grid

Protein Design

Given a three-dimensional structure, determine a sequence of amino acids for a protein that would due to their combined bio-chemical properties assume that three-dimensional structure.
Protein Design on the Illinois Bio-Grid

Protein Design
- NP-Hard
- Brute Force Search is computationally infeasible

- Any solution will require
  - Models using biologically relevant information
  - Algorithms to search the enormous problem space
Protein Design on the Illinois Bio-Grid

Protein Data Bank
- 39,323 structures (and counting)
- Calculate propensities in the data
- Why try and solve a problem that nature has already figured out?

Phi Angles  Psi Angles  Environment  Orientation  Frequency  Nearest Neighbors
0.000000  0.185175  0.094890  0.000000  0.000000
0.033546  0.084551  0.032626  0.028567  0.000000
0.000000  0.000000  0.070034  0.107855  0.000000
0.000000  0.044409  0.157549  0.135519  0.025277

Phi Angles  Psi Angles  Environment  Orientation  Frequency  Nearest Neighbors
0.000000  0.185175  0.094890  0.000000  0.000000
0.033546  0.084551  0.032626  0.028567  0.000000
0.000000  0.000000  0.070034  0.107855  0.000000
0.000000  0.044409  0.157549  0.135519  0.025277
Protein Design on the Illinois Bio-Grid

Incorporate the knowledge gleaned from the PDB into algorithms capable of determining the correct amino acid sequence.

- Monte Carlo
  Randomly sample the problem space, zeroing in on potential solutions

- K-Nearest Neighbor
  Look for similar instances in nature

- Other Machine Learning and Data Analysis Techniques
Protein Design on the Illinois Bio-Grid

Once we determine a sequence, it is still hard to know if we are right.

- Computationally fold the sequence
  - Also NP-Hard
- Actually build the protein in the wet lab and examine the structure
  - Expensive
  - Time-consuming
  - May be impossible
- Trust secondary evidence
  - Not necessarily reliable
Protein Design on the Illinois Bio-Grid

However, despite the difficulty, the rewards from success could be incredible.

- Compstatin, prevents the autoimmune-mediated damage of organs during transplantation, and various inflammatory diseases.
Protein Design on the Illinois Bio-Grid

Other Illinois Bio-Grid Research Projects

- Bioinformatics Interactive Programming Environment (Pelagic)
- High Throughput Task Allocator
- Gene Designer
- Unfolded State Predictor
- Rama Map
- Mass Spectrum Analysis

Contact Prof. Angulo for more information about the IBG.  dangulo@cti.depaul.edu
Protein Design on the Illinois Bio-Grid

This work was supported in part by the National Science Foundation under Grant No. 0353989.