Computational Modeling of Molecular Structure

Jianlin Cheng, PhD
Computer Science Department
Informatics Institute
University of Missouri, Columbia
Spring, 2014
The Genomic Era

Collins, Venter, Human Genome, 2000
DNA Sequencing Revolution

$1000 Personal Genome by 2020

Scientists

Government

Company
A Topic of Big Bio Data Analysis

Science enters $1,000 genome era

By Paul Rincon
Science editor, BBC News website

The HiSeq X Ten is capable of sequencing five human genomes a day, Illumina claims
Objectives

• Properties of molecular structures (proteins, RNA, genome / DNA)
• Computational representation of molecular structures
• Computational modeling of molecular structures
• Application of modeling of molecular structures
Significance of Studying Molecular Structures

- One foundation of life sciences
- Personal healthcare and medicine
- One major topic of bioinformatics and computational biology – an important field of computer science
- A great application area of computer algorithms and data structures
- A great application area of engineering
- A very interdisciplinary field (CS, math, biology, chemistry, physics)
Three Kinds of Structures

- Protein Structure
- Genome Structure
- RNA Structure
Representation of Molecular Structures

- X, Y, Z coordinates
- Euclidean grid
- Vector and angles
- Computer graphics
Algorithms

- Grid-based simulation (random walk)
- Vector-based simulation
- Angular-based simulation
- Gradient descent simulation and variants
- Simulated annealing
- Markov Chain Monte Carlo
- Probabilistic modeling
- Constraint-based optimization
Software Packages

• RasMol, Jmol, PyMol
• Modeller, Rosetta, I-TASSER, MULTICOM, CNS, etc
• Your own algorithm, implementation, and practice
Course Format

• Course web site: 
  https://people.cs.missouri.edu/~chengji/cscmms/

• Problem solving

• Active learning by practicing

• Syllabus (see details)
Teaching Format of Each Topic

Course Introduction

Topic Lecture (reading)

Problem Definition (discussion, planning)

Plan Presentation

Project Implementation (programming)

Results and Analysis (report and update)

Final presentation and report

Group:

5 students per group

Rotate as topic coordinator

Each member participates in every topic

All members present the whole project
Grading

• Class discussions (15%)
• Literature reviews (10%)
• Topic plan presentation (20%, group)
• Topic implementation and report (45%, group)
• Final presentation (10%, group)
Introduction to Molecular Biology for Computer Science and Engineering Students
Introduction to Molecular Biology

- Cell is the unit of structure and function of all living things.

Two types of cells: eukaryote (higher organisms) and prokaryote (lower organisms)
Central Dogma of Molecular Biology

- DNA
- RNA
- Protein

Replication → Transcription → Translation → Phenotype

Genotype
Genes contain instructions for making proteins. Proteins act alone or in complexes to perform many cellular functions.
Central Dogma of Molecular Biology

Information flow

Replication

DNA

RNA

Protein

Transcription

Translation

Reverse Transcription (HIV virus)
DNA (Deoxyribose Nucleotide Acids)

DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a "polynucleotide." Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group.

A is for adenine  
G is for guanine  
C is for cytosine  
T is for thymine

CGAATGGGGAAA......

Introduction to DNA structure, Richard B. Hallick, 1995
Base Pairs:
A-T (2 H-bonds)
C-G (3 H-bonds)

Hydrogen bonds: non-covalent bonds mediated by hydrogen atoms
Uncoiled DNA Molecule

Source: Dr. Gary Stormo, 2002
Fundamental Problems: How genetic information pass from one cell to another and from one generation to next generation
DNA Replication

DNA Polymerase
RNA (Ribose Nucleotide Acids)

ACGAAUAACAGGUAUAUAAAAAUAGAUAUUACCUAUAGAUUUCGU
Different Kinds of RNA

• mRNA: messager RNA
carry genetic information out of nucleus for protein synthesis
(transcription process: RNA polymerase)
• rRNA: ribosomal RNA
constitute 50% of ribosome, which is a molecular assembly for
protein synthesis
• tRNA: transfer RNA
decode information (map 3 nucleotides to amino acid);
transfer amino acid
• snRNA: small RNA molecules found in nucleus
involve RNA splicing
• Non-coding RNA
Transcription of Gene into RNA
Three Nucleotides is called a codon.
Protein Sequence

A directional sequence of amino acids/residues

H₂N─CH─C─NH─CH─C─NH─CH─C─NH─CH─C─NH─CH─C─...
Amino Acid Structure
Lysine
## Amino Acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Abbrev.</th>
<th>Side Chain</th>
<th>Hydrophobic</th>
<th>Polar</th>
<th>Charged</th>
<th>Small</th>
<th>Tiny</th>
<th>Aromatic or Aliphatic</th>
<th>van der Waals Volume</th>
<th>Codon</th>
<th>Occurrence in proteins (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala, A</td>
<td>-CH₃</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>Aromatic or Aliphatic</td>
<td>67</td>
<td>GCU, GCC, GCA, GCG</td>
<td>7.8</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys, C</td>
<td>-CH₂SH</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>Aromatic or Aliphatic</td>
<td>86</td>
<td>UGU, UGC</td>
<td>6.9</td>
</tr>
<tr>
<td>Aspartate</td>
<td>Asp, D</td>
<td>-CH₂COOH</td>
<td>-</td>
<td>X</td>
<td>negative</td>
<td>X</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>91</td>
<td>GAA, GAG</td>
<td>5.3</td>
</tr>
<tr>
<td>Glutamate</td>
<td>Glu, E</td>
<td>-CH₂CH₂COOH</td>
<td>-</td>
<td>X</td>
<td>negative</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>109</td>
<td>GAA, GAG</td>
<td>6.3</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe, F</td>
<td>-CH₂C₆H₅</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>135</td>
<td>UUU, UUC</td>
<td>3.9</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly, G</td>
<td>-H</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>Aromatic or Aliphatic</td>
<td>48</td>
<td>GGU, GGC, GGA, GGG</td>
<td>7.2</td>
</tr>
<tr>
<td>Histidine</td>
<td>His, H</td>
<td>-CH₃C₆H₄N₂</td>
<td>-</td>
<td>X</td>
<td>positive</td>
<td>X</td>
<td>X</td>
<td>Aromatic or Aliphatic</td>
<td>118</td>
<td>CAU, CAC</td>
<td>2.3</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile, I</td>
<td>-CH(CH₃)₂</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>124</td>
<td>AUU, AUC, AUA</td>
<td>5.3</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys, K</td>
<td>-(CH₂)₄NH₂</td>
<td>-</td>
<td>X</td>
<td>positive</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>135</td>
<td>AAA, AAG</td>
<td>5.9</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu, L</td>
<td>-CH₂CH(CH₃)₂</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>124</td>
<td>UUA, UUG, CUU, CUC, CUA, CUG</td>
<td>9.1</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met, M</td>
<td>-CH₂CH₂SCH₃</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>124</td>
<td>AUG</td>
<td>2.3</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn, N</td>
<td>-CH₂CONH₂</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>96</td>
<td>AAU, AAC</td>
<td>4.3</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro, P</td>
<td>-CH₂CH₂CH₂</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>90</td>
<td>CCU, CCC, CCA, CCG</td>
<td>5.2</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln, Q</td>
<td>-CH₂CH₂CONH₂</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>114</td>
<td>CAA, CAG</td>
<td>4.2</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg, R</td>
<td>(CH₂)₃NH-[O(NH)NH₂]</td>
<td>-</td>
<td>X</td>
<td>positive</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>148</td>
<td>CGU, CBC, CGA, CGG, AGA, AGG</td>
<td>5.1</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser, S</td>
<td>-CH₂OH</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>Aromatic or Aliphatic</td>
<td>73</td>
<td>UCU, UCC, UCA, UCG, AGU, AGC</td>
<td>6.8</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr, T</td>
<td>-CH(OH)CH₃</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>93</td>
<td>ACC, ACA, ACG</td>
<td>5.9</td>
</tr>
<tr>
<td>Valine</td>
<td>Val, V</td>
<td>-CH(CH₃)₂</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>105</td>
<td>GUU, GUC, GUA, GUG</td>
<td>6.6</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp, W</td>
<td>-CH₂C₆H₅N</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>163</td>
<td>UGG</td>
<td>1.4</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr, Y</td>
<td>-CH₂C₆H₄OH</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Aromatic or Aliphatic</td>
<td>141</td>
<td>LIAU, UAC</td>
<td>3.2</td>
</tr>
</tbody>
</table>

**Hydrophilic**
Central Dogma of Proteomics

AGCWY......

Sequence Structure Function

Cell
The Genomic Era

Collins, Venter, Human Genome, 2000
Personal Genome’s Implications

- Personalized Disease Prevention
- Personalized Disease Diagnosis
- Personalized Medicine
- Personalized Health Care
Genome Implications to Information Sciences and Life Sciences

Elements and Systems
Assignment One


Submit your review summary to mudatamining@gmail.com. Due by Feb. 3 (Monday).
images.google.com and all the authors providing valuable images