“Bioinformatics is the application of computers to the collection, archiving, organization, and analysis of biological data.”

... A hybrid of biology and computer science

“Bioinformatics is computer aided biology!

Bioinformatics is computer aided biology!

Goal: Data to Knowledge
So what is **structural bioinformatics**?

Aims to characterize and interpret biomolecules and their assemblies at the molecular & atomic level

---

Why should we care?

Because biomolecules are “nature’s robots”

... and because it is only by coiling into **specific 3D structures** that they are able to perform their functions
BIOINFORMATICS DATA

Genomes
DNA & RNA sequence
DNA & RNA structure
Protein sequence
Protein families, motifs and domains
Protein structure
Protein interactions
Chemical entities
Pathways
Systems
Gene expression
Literature and ontologies

STRUCTURAL DATA IS CENTRAL

Genomes
DNA & RNA sequence
DNA & RNA structure
Protein sequence
Protein families, motifs and domains
Protein structure
Protein interactions
Chemical entities
Pathways
Systems
Gene expression
Literature and ontologies

Sequence > Structure > Function

ENERGETICS
DYNAMICS

Sequence & Structure & Function
In daily life, we use machines with functional **structure** and **moving parts**

Genomics is a great start ....

- But a parts list is not enough to understand how a bicycle works

  ---

  **Track Bike — DL 175**

<table>
<thead>
<tr>
<th>NO.</th>
<th>IH No.</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>156515</td>
<td>Track Frame 82° 26&quot; W/ 20&quot; 28&quot; 34&quot; Team Red</td>
</tr>
<tr>
<td>2</td>
<td>157280</td>
<td>Fork For 28&quot; Frame</td>
</tr>
<tr>
<td>3</td>
<td>157297</td>
<td>Fork For 20&quot; Frame</td>
</tr>
<tr>
<td>4</td>
<td>157303</td>
<td>Fork For 20&quot; Frame</td>
</tr>
<tr>
<td>5</td>
<td>157307</td>
<td>Fork For 24&quot; Frame</td>
</tr>
<tr>
<td>6</td>
<td>151200</td>
<td>Handlebar TTT Competition Track Alloy 15/16&quot;</td>
</tr>
<tr>
<td>7</td>
<td>151379</td>
<td>Handlebar Stem TTT: specific extension</td>
</tr>
<tr>
<td>8</td>
<td>151377</td>
<td>Cromoly Bell</td>
</tr>
<tr>
<td>9</td>
<td>153462</td>
<td>Nonskid Complete 1 x 24 Disc</td>
</tr>
<tr>
<td>10</td>
<td>153464</td>
<td>Belt Drivetrain</td>
</tr>
<tr>
<td>11</td>
<td>150402</td>
<td>175 Rolando Fostated Disk Tubular Presta Sake 75&quot;</td>
</tr>
<tr>
<td>12</td>
<td>150423</td>
<td>Eax 27&quot; W/ ARA Competition 1300 Alus PrestaSake</td>
</tr>
<tr>
<td>13</td>
<td>150540</td>
<td>Chain 1/4 Chain</td>
</tr>
<tr>
<td>14</td>
<td>150539</td>
<td>Gears 1/2/3/4 Speed</td>
</tr>
<tr>
<td>15</td>
<td>150537</td>
<td>Brake Levers</td>
</tr>
<tr>
<td>16</td>
<td>150538</td>
<td>Brake Levers</td>
</tr>
<tr>
<td>17</td>
<td>150557</td>
<td>Brake Levers</td>
</tr>
<tr>
<td>18</td>
<td>150558</td>
<td>Brake Levers</td>
</tr>
<tr>
<td>19</td>
<td>150559</td>
<td>Brake Levers</td>
</tr>
<tr>
<td>20</td>
<td>150560</td>
<td>Brake Levers</td>
</tr>
<tr>
<td>21</td>
<td>151372</td>
<td>Freewheel Nut</td>
</tr>
<tr>
<td>22</td>
<td>151373</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>23</td>
<td>151374</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>24</td>
<td>151375</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>25</td>
<td>151376</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>26</td>
<td>151377</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>27</td>
<td>151378</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>28</td>
<td>151379</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>29</td>
<td>151380</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>30</td>
<td>151381</td>
<td>Freewheel Hub</td>
</tr>
<tr>
<td>31</td>
<td>151382</td>
<td>Freewheel Hub</td>
</tr>
</tbody>
</table>

... but not the end

- We want the full spatiotemporal picture, and an ability to control it
- Broad applications, including drug design, medical diagnostics, chemical manufacturing, and energy
KEY CONCEPT: ENERGY LANDSCAPE

**Sequence**
- Unfolded chain of amino acid chain
- Highly mobile
- Inactive

**Structure**
- Ordered in a precise 3D arrangement
- Stable but dynamic

**Function**
- Active in specific “conformations”
- Specific associations & precise reactions

Barrier crossing time \(\sim \exp(\text{Barrier Height})\)

0.1 microseconds

1 millisecond

Native
Compact, Ordered

Unfolded
Expanded, Disordered

Molten Globule
Compact, Disordered

KEY CONCEPT: ENERGY LANDSCAPE
Today’s Menu

- Overview of structural bioinformatics
  - Motivations, goals and challenges
- Fundamentals of protein structure
  - Structure composition, form and forces
- Representing, interpreting & modeling protein structure
  - Visualizing & interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

TRADITIONAL FOCUS PROTEIN, DNA AND SMALL MOLECULE DATA SETS WITH MOLECULAR STRUCTURE

Protein (PDB)  DNA (NDB)  Small Molecules (CCDB)
**PDB – A Billion Atom Archive**

> 1 billion atoms in the asymmetric units

1,000,000 Structures in May 2018

~151,500 Structures as of May 2019

**Motivation 1:**
Detailed understanding of molecular interactions

Provides an invaluable structural context for conservation and mechanistic analysis leading to functional insight.

**Motivation 2:**
Lots of structural data is becoming available

Structural Genomics has contributed to driving down the cost and time required for structural determination

Data from: [https://www.rcsb.org/stats/](https://www.rcsb.org/stats/)

---

Motivation 2:
Lots of structural data is becoming available

Structural Genomics has contributed to driving down the cost and time required for structural determination

Image Credit: “Structure determination assembly line” Adam Godzik

Motivation 3:
Theoretical and computational predictions have been, and continue to be, enormously valuable and influential!

SUMMARY OF KEY MOTIVATIONS

Sequence > Structure > Function
• Structure determines function, so understanding structure helps our understanding of function

Structure is more conserved than sequence
• Structure allows identification of more distant evolutionary relationships

Structure is encoded in sequence
• Understanding the determinants of structure allows design and manipulation of proteins for industrial and medical advantage

Goals:
• Visualization
• Analysis
• Comparison
• Prediction
• Design

Goals:

- Visualization
- Analysis
- Comparison
- Prediction
- Design


Goals:

- Visualization
- Analysis
- Comparison
- Prediction
- Design

Grant et al. unpublished

Goals:

- Visualization
- Analysis
- Comparison
- Prediction
- Design


Goals:

- Visualization
- Analysis
- Comparison
- Prediction
- Design

Grant et al. PLoS Biology (2011)
MAJOR RESEARCH AREAS AND CHALLENGES

Include but are not limited to:

• Protein classification
• Structure prediction from sequence
• Binding site detection
• Binding prediction and drug design
• Modeling molecular motions
• Predicting physical properties (stability, binding affinities)
• Design of structure and function
• etc...

With applications to Biology, Medicine, Agriculture and Industry

Today’s Menu

• Overview of structural bioinformatics
  • Motivations, goals and challenges

  • Fundamentals of protein structure
    • Structure composition, form and forces

  • Representing, interpreting & modeling protein structure
    • Visualizing & interpreting protein structures
    • Analyzing protein structures
    • Modeling energy as a function of structure

HIERARCHICAL STRUCTURE OF PROTEINS

Primary > Secondary > Tertiary > Quaternary

ReCAP: AMINO ACID NOMENCLATURE
AMINO ACIDS CAN BE GROUPED BY THE PHYSIOCHEMICAL PROPERTIES

AMINO ACIDS POLYMERIZE THROUGH PEPTIDE BOND FORMATION

AMINO ACIDS POLYMERIZE THROUGH PEPTIDE BOND FORMATION

PEPTIDES CAN ADOPT DIFFERENT CONFORMATIONS BY VARYING THEIR PHI & PSI BACKBONE TORSIONS

PHI VS PSI PLOTS ARE KNOWN AS RAMACHANDRAN DIAGRAMS

- Steric hindrance dictates torsion angle preference
- Ramachandran plot show preferred regions of $\phi$ and $\psi$ dihedral angles which correspond to major forms of secondary structure
**MAJOR SECONDARY STRUCTURE TYPES**

**ALPHA HELIX & BETA SHEET**

### Alpha Helix
- Most common from has **3.6 residues per turn** (number of residues in one full rotation)
- Hydrogen bonds (dashed lines) between residue \(i\) and \(i+4\) stabilize the structure
- The side chains (in green) protrude outward
- **\(3_{10}\)-helix** and **\(\pi\)-helix** forms are less common


### Beta Sheet
- **In antiparallel \(\beta\)-sheets**
  - Adjacent \(\beta\)-strands run in **opposite** directions
  - Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
  - The side chains (in green) are above and below the sheet


- **In parallel \(\beta\)-sheets**
  - Adjacent \(\beta\)-strands run in **same** direction
  - Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
  - The side chains (in green) are above and below the sheet


---

**What Does a Protein Look like?**
• Proteins are stable (and hidden) in water

• Proteins closely interact with water

• Proteins are close packed solid but flexible objects (globular)

• Due to their large size and complexity it is often hard to see what’s important in the structure
• Backbone or main-chain representation can help trace chain topology

• Simplified secondary structure representations are commonly used to communicate structural details
  • Now we can clearly see $2^\circ$, $3^\circ$, and $4^\circ$ structure
  • Coiled chain of connected secondary structures

• Backbone or main-chain representation can help trace chain topology & reveal secondary structure

DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY

Superposition of all 482 structures in RCSB PDB (23/09/2015)
**DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY**

**Key Concept:** **ENERGY LANDSCAPE**

- **Native State(s):** Compact, Ordered
- **Molten Globule State:** Unfolded
- **Expanded, Disordered
- **Compact, Disordered**

- **Barrier crossing time:** ~$\text{exp}(\text{Barrier Height})$
- **Barrier Height:** 0.1 microseconds
- **1 millisecond**
- **Multiple Native Conformations** (e.g. ligand bound and unbound)

**Normal Mode Analysis (NMA)** models the protein as a network of elastic strings.

**Key forces affecting structure:**

- **H-bonding**
- **Van der Waals**
- **Electrostatics**
- **Hydrophobicity**

- **Hydrogen-bond donor**
- **Hydrogen-bond acceptor**

- $N-H-N$
- $N-H-O$
- $O-H-N$
- $O-H-O$

- $2.6 \text{ Å} < d < 3.1 \text{ Å}$
- $150^\circ < \theta < 180^\circ$
Key forces affecting structure:

- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity

\[ \Delta E = \frac{A}{r^2} - \frac{B}{r^6} \]

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>( \delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>3 Å &lt; d &lt; 4 Å</td>
</tr>
</tbody>
</table>

The force that causes hydrophobic molecules or nonpolar portions of molecules to aggregate together rather than to dissolve in water is called Hydrophobicity (Greek, “water fearing”). This is not a separate bonding force; rather, it is the result of the energy required to insert a nonpolar molecule into water.

Key forces affecting structure:

- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity

- (some time called IONIC BONDS or SALT BRIDGES)

Key forces affecting structure:

- H-bonding
- Van der Waals
- Electrostatics
- Hydrophobicity

\[ E = K \frac{q_1 q_2}{D r} \]

Coulomb’s law

- \( E \) = Energy
- \( k \) = constant
- \( D \) = Dielectric constant (vacuum = 1; \( H_2O = 80 \))
- \( q_1 \) & \( q_2 \) = electronic charges (Coulombs)
- \( r \) = distance (Å)

Today’s Menu

- Overview of structural bioinformatics
- Motivations, goals and challenges
- Fundamentals of protein structure
  - Structure composition, form and forces
- Representing, interpreting & modeling protein structure
  - Visualizing & interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure
Today’s Menu

• Overview of structural bioinformatics
  • Motivations, goals and challenges

• Fundamentals of protein structure
  • Structure composition, form and forces

• Representing, interpreting & modeling protein structure
  • Visualizing & interpreting protein structures
  • Analyzing protein structures
  • Modeling energy as a function of structure

Do it Yourself!
Hand-on time!

https://bioboot.github.io/bggn213_S19/lectures/#11

Focus on section 1 only please!

N.B. Remember to make your new class11 RStudio project inside your GitHub tracked directory from last day and UNCHECK the "Create a Git repository" option...

SIDE-NOTE: PDB FILE FORMAT

• PDB files contains atomic coordinates and associated information.
Today’s Menu

- Overview of structural bioinformatics
  - Motivations, goals and challenges
- Fundamentals of protein structure
  - Structure composition, form and forces
- Representing, interpreting & modeling protein structure
  - Visualizing and interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

Hand-on time!

Focus on section 2 please!

N.B. You will need to have VMD installed on your computer (see class website and hands-on sheet for details)

Hand-on time!

Focus on section 3 to 5

Hand-on time!

Focus on section 6

Working with Multiple Structures and large structure ensembles from experiment and theory
Download MUSCLE for your OS from:
https://www.drive5.com/muscle/downloads.htm

On MAC use your TERMINAL to enter the commands:
```
tar -xvf ~/Downloads/muscle3.8.31_i86darwin32.tar
sudo mv muscle3.8.31_i86darwin32 /usr/local/bin/muscle
```

On Windows use file explorer to:
- Move the downloaded muscle3.8.31_i86win32.exe from your Downloads folder to your Project folder.
- Then right click to rename to muscle.exe

To use in your R session:
```
library("bio3d.view")
pdb <- read.pdb("5p21")
view(pdb)
view(pdb, "overview", col="sse")
```

Today’s Menu

- Overview of structural bioinformatics
  - Motivations, goals and challenges

- Fundamentals of protein structure
  - Structure composition, form and forces

- Representing, interpreting & modeling protein structure
  - Visualizing and interpreting protein structures
  - Analyzing protein structures
  - Modeling energy as a function of structure

Bio3D view()

- If you want the 3D viewer in your R markdown you can install the development version of bio3d.view

```
install.packages("devtools")
devtools::install_bitbucket("Grantlab/bio3d-view")
```

NMA in Bio3D

- Normal Mode Analysis (NMA) is a bioinformatics method that can predict the major motions of biomolecules.

```
library(bio3d)
library(bio3d.view)
```

```
pdb <- read.pdb("1hel")
modes <- nma(pdb)
m7 <- mktrj(modes, mode=7, file="mode_7.pdb")
view(m7, col=vec2color(rmsf(m7)))
```
SideNote: `view()`

- If you want the interactive 3D viewer in Rmd rendered to output: html_output document:

```r
library(bio3d.view)
library(rgl)
```

```r
modes <- nma(read.pdb("1hel"))
m7 <- mktrj(modes, mode=7, file="mode_7.pdb")
view(m7, col=vec2color(rmsf(m7)))
rglwidget(width=500, height=500)
```

Optional:
Stop here for Today!

[ Muddy Point Assessment ]

SUMMARY

- Structural bioinformatics is computer aided structural biology
- Described major motivations, goals and challenges of structural bioinformatics
- Reviewed the fundamentals of protein structure
- Explored how to use R to perform advanced custom structural bioinformatics analysis!
- Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally

[ Muddy Point Assessment ]