MODULE OVERVIEW

Objective: Provide an introduction to the practice of bioinformatics as well as a practical guide to using common bioinformatics databases and algorithms

1.1. Introduction to Bioinformatics
1.2. Sequence Alignment and Database Searching
1.3. Structural Bioinformatics
1.4. Genome Informatics: High Throughput Sequencing Applications and Analytical Methods

WEEK TWO REVIEW

- Answers to last week’s homework (19/19):
  - Answers week 2
- Muddy Point Assessment (11/19):
  - Responses
    - “More time to finish the assignment”
    - “I felt there was too much material to cover in one lab”
    - “The [NCBI] sites were so slow”
    - “More time with HMMER would be helpful”
    - “Very nice lab”

Q18: NW DYNAMIC PROGRAMMING

Match: +2
Mismatch: -1
Gap: -2

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-2</td>
<td>-4</td>
<td>-6</td>
<td>-8</td>
<td>-10</td>
</tr>
<tr>
<td>A</td>
<td>-2</td>
<td>0</td>
<td>-2</td>
<td>-4</td>
<td>-6</td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td>0</td>
<td>+1</td>
<td>+2</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>-6</td>
<td>-2</td>
<td>+3</td>
<td>+4</td>
<td>+2</td>
</tr>
<tr>
<td>G</td>
<td>-8</td>
<td>-4</td>
<td>0</td>
<td>+1</td>
<td>+2</td>
</tr>
<tr>
<td>C</td>
<td>-10</td>
<td>-6</td>
<td>-2</td>
<td>+1</td>
<td>0</td>
</tr>
</tbody>
</table>

A - TT GC
I I I I
A G T T C

A - TT GC
I I I I
A G T T - C

Check out the “Background Reading” material online:
- Achievements & Challenges in Structural Bioinformatics
- Protein Structure Prediction
- Biomolecular Simulation
- Computational Drug Discovery

Complete the lecture 1.3 homework questions:
http://tinyurl.com/bioinf525-quiz3

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

Bioinformatics is computer aided biology!

So what is structural bioinformatics?

... computer aided structural biology!

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

**ENERGETICS**

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

**DYNAMICS**

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

**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

- 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**

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

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

- Active in specific “conformations”
- Specific associations & precise reactions
KEY CONCEPT: ENERGY LANDSCAPE

Multiple Native Conformations (e.g., ligand bound and unbound)

OUTLINE:

- Overview of structural bioinformatics
  - Major motivations, goals and challenges
- Fundamentals of protein structure
  - Composition, form, forces and dynamics
- Representing and interpreting protein structure
  - Modeling energy as a function of structure
- Example application areas
  - Predicting functional dynamics & drug discovery

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

- Protein (PDB)
- DNA (NDB)
- Small Molecules (CCDB)

Motivation 1:
Detailed understanding of molecular interactions

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

Motivation 1:
Detailed understanding of molecular interactions

Computational modeling can provide detailed insight into functional interactions, their regulation and potential consequences of perturbation.

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: http://www.rcsb.org/pdb/statistics/

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:
- Analysis
- Visualization
- Comparison
- Prediction
- Design

Grant et al. JMB. (2007)

Goals:
- Analysis
- Visualization
- Comparison
- Prediction
- Design

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

NEXT UP:

• Overview of structural bioinformatics
  • Major motivations, goals and challenges

  • Fundamentals of protein structure
    • Composition, form, forces and dynamics

  • Representing and interpreting protein structure
    • Modeling energy as a function of structure

  • Example application areas
    • Predicting functional dynamics & drug discovery
**Hierarchical Structure of Proteins**

- Primary
- Secondary
- Tertiary
- Quaternary

**Amino Acid Residues**
- Alpha helix
- Polypeptide chain
- Assembled subunits

**Recap: Amino Acid Nomenclature**

- Main chain (backbone)
- Side chain (R group)

**Amino Acids Can Be Grouped by the Physiochemical Properties**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leucine</td>
<td>Nonpolar, aliphatic</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Polar, aromatic</td>
</tr>
<tr>
<td>Lysine</td>
<td>Positive charge</td>
</tr>
<tr>
<td>Glycine</td>
<td>Nonpolar, small</td>
</tr>
</tbody>
</table>

**Amino Acids Polymerize Through Peptide Bond Formation**

- Side chains
- Backbone

**Peptides Can Adopt Different Conformations by Varying Their Phi & Psi Backbone Torsions**

- Bond angles and lengths largely invariant
- Peptide bond is planar (Cα, C, O, N, H, Cα all lie in the same plane)

**Phi vs Psi Plots Are Known as Ramachandran Diagrams**

- Steric hindrance dictates torsion angle preference
- Ramachandran plot shows preferred regions of ϕ and ψ dihedral angles which correspond to major forms of secondary structure
**MAJOR SECONDARY STRUCTURE TYPES**

**ALPHA HELIX & BETA SHEET**

**α-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 π-helix forms are less common

Hydrogen bond: i→i+4


**In antiparallel β-sheets**
- Adjacent β-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 β-sheets**
- Adjacent β-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

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

• Simplified secondary structure representations are commonly used to communicate structural details
• Now we can clearly see 2nd, 3rd and 4th structure
• Coiled chain of connected secondary structures

DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY

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

Principal component analysis (PCA) of experimental structures

Key forces affecting structure:
• H-bonding
• Van der Waals
• Electrostatics
• Hydrophobicity
• Disulfide Bridges

Key forces affecting structure:
• H-bonding
• Van der Waals
• Electrostatics
• Hydrophobicity
• Disulfide Bridges

Key forces affecting structure:
• H-bonding
• Van der Waals
• Electrostatics
• Hydrophobicity
• Disulfide Bridges

Key forces affecting structure:
• H-bonding
• Van der Waals
• Electrostatics
• Hydrophobicity
• Disulfide Bridges

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.

**Coulomb's law**

\[
E = \frac{K q_1 q_2}{d^2}
\]

- \(E\) = Energy
- \(K\) = constant
- \(D\) = Dielectric constant (vacuum = 1; H\(_2\)O = 80)
- \(q_1\) & \(q_2\) = electronic charges (Coulombs)
- \(d\) = distance (Å)
Forces affecting structure:

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

Other names:
cystine bridge
disulfide bridge

Hair contains lots of disulfide bonds which are broken and reformed by heat.

NEXT UP:

- Overview of structural bioinformatics
  - Major motivations, goals and challenges
- Fundamentals of protein structure
  - Composition, form, forces and dynamics
- Representing and interpreting protein structure
  - Modeling energy as a function of structure
- Example application areas
  - Predicting functional dynamics & drug discovery

Search: HIV

Search: 1HSG (PDB ID)
PDB FILE FORMAT

- PDB files contain atomic coordinates and associated information.

KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEM'S ENERGY AS A FUNCTION OF ITS STRUCTURE

Two main approaches:
1. Physics-Based
2. Knowledge-Based

PHYSICS-BASED POTENTIALS
ENERGY TERMS FROM PHYSICAL THEORY

CHARMM P.E. function, see: http://www.charmm.org/

TOTAL POTENTIAL ENERGY

- The total potential energy or enthalpy fully defines the system, U.
- The forces are the gradients of the energy.
- The energy is a sum of independent terms for:
  - Bonds
  - Bond angles
  - Torsion angles
  - Non-bonded atom pairs

MOVING OVER THE ENERGY SURFACE

- Energy Minimization drops into local minimum.
- Molecular Dynamics uses thermal energy to move smoothly over surface.
- Monte Carlo Moves are random. Accept with probability exp(-ΔU/kT).

Slide Credit: Michael Levitt
PHYSICS-ORIENTED APPROACHES

Weaknesses
Fully physical detail becomes computationally intractable
Approximations are unavoidable
(Quantum effects approximated classically, water may be treated crudely)
Parameterization still required

Strengths
Interpretable, provides guides to design
Broadly applicable, in principle at least
Clear pathways to improving accuracy

Status
Useful, widely adopted but far from perfect
Multiple groups working on fewer, better approx
Force fields, quantum
entropy, water effects
Moore’s law: hardware improving

SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER

KEY CONCEPT: POTENTIAL FUNCTIONS DESCRIBE A SYSTEM’S ENERGY AS A FUNCTION OF ITS STRUCTURE

Two main approaches:
(1). Physics-Based
(2). Knowledge-Based

KNOWLEDGE-BASED DOCKING POTENTIALS
ENERGY DETERMINES PROBABILITY (STABILITY)
Basic idea: Use probability as a proxy for energy

\[ p(r) \propto e^{-E(r)/kT} \]
Inverse Boltzmann:
\[ E(r) = -kT \ln[p(r)] \]

Example: ligand carboxylate O to protein histidine N
Find all protein-ligand structures in the PDB with a ligand carboxylate O
1. For each structure, histogram the distances from O to every histidine N
2. Sum the histograms over all structures to obtain \( p(\text{O-N}) \)
3. Compute \( E(\text{O-N}) \) from \( p(\text{O-N}) \)

KNOWLEDGE-BASED DOCKING POTENTIALS

A few types of atom pairs, out of several hundred total

\[ E_{\text{prot-lig}} = E_{\text{obs}} + \sum_{\text{pair}(i,j)} E_{\text{app}(i,j)}(r_{ij}) \]

KNOWLEDGE-BASED POTENTIALS
Weaknesses
Accuracy limited by availability of data

Strengths
Relatively easy to implement
Computationally fast

Status
Useful, far from perfect
May be at point of diminishing returns
(not always clear how to make improvements)

NEXT UP:
› Overview of structural bioinformatics
  • Major motivations, goals and challenges
› Fundamentals of protein structure
  • Composition, form, forces and dynamics
› Representing and interpreting protein structure
  • Modeling energy as a function of structure

Example application areas
• Predicting functional dynamics & drug discovery

PREDICTING FUNCTIONAL DYNAMICS
• Proteins are intrinsically flexible molecules with internal motions that are often intimately coupled to their biochemical function
  – E.g. ligand and substrate binding, conformational activation, allosteric regulation, etc.

• Thus knowledge of dynamics can provide a deeper understanding of the mapping of structure to function
  – Molecular dynamics (MD) and normal mode analysis (NMA) are two major methods for predicting and characterizing molecular motions and their properties

MOLECULAR DYNAMICS SIMULATION
• Use force-field to find
Potential energy between all atom pairs
• Move atoms to next state
• Repeat to generate trajectory

McCammon, Gelin & Karplus, Nature (1977)
[ See: https://www.youtube.com/watch?v=u1ZyeMfK3 ]
Divide time into discrete (~1fs) time steps (\( \Delta t \)) (for integrating equations of motion, see below)

At each time step calculate pair-wise atomic forces \( (F(t)) \) (by evaluating force-field gradient)

Use the forces to calculate velocities and move atoms to new positions (by integrating numerically via the "leapfrog" scheme)

REPEAT, (iterate many, many times...1ms = 10^{12} time steps)

**BASIC ANATOMY OF A MD SIMULATION**

**MD Prediction of Functional Motions**

Yao and Grant, Biophys J. (2013)

**Simulations Identify Key Residues Mediating Dynamic Activation**

Yao ... Grant, Journal of Biological Chemistry (2016)
EXAMPLE APPLICATION OF MOLECULAR SIMULATIONS TO GPCRS

Binding

GPCR

G protein

Cell

Membrane

COARSE GRAINING: NORMAL MODE ANALYSIS (NMA)

- MD is still time-consuming for large systems
- Elastic network model NMA (ENM-NMA) is an example of a lower resolution approach that finishes in seconds even for large systems.

ATOMISTIC

Coarse Grained

• 1 bead / 1 amino acid
• Connected by springs

COARSE GRAINING:
NORMAL MODE ANALYSIS (NMA)

• Example: F$_2$-ATPase in water (183,674 atoms) for 1 nanosecond:
  => 10^8 integration steps
  => 8.4 * 10^{11} floating point operations/step
  \[ n(n-1)/2 \] interactions

  Total: 8.4 * 10^{12} flop
  (on a 100 Gflop/s cpu: \textbf{ca. 25 years})

... but performance has been improved by use of:
- multiple time stepping \textbf{ca. 2.5 years}
- fast multipole methods \textbf{ca. 1 year}
- parallel computers \textbf{ca. 5 days}
- modern GPUs \textbf{ca. 1 day}
- modern GPUs (Anton supercomputer) \textbf{ca. minutes}

NMA MODELS THE PROTEIN AS A NETWORK OF ELASTIC STRINGS

Proteinase K

NEXT UP:

* Overview of structural bioinformatics
  * Major motivations, goals and challenges
* Fundamentals of protein structure
  * Composition, form, forces and dynamics
* Representing and interpreting protein structure
  * Modeling energy as a function of structure
* Example application areas
  * Predicting functional dynamics & drug discovery
THE TRADITIONAL EMPIRICAL PATH TO DRUG DISCOVERY

- Compound library (commercial, in-house, synthetic, natural)
- High throughput screening (HTS)
- Hit confirmation
- Lead compounds (e.g., pM $K_d$)
- Lead optimization (Medicinal chemistry)
- Potent drug candidates (nM $K_d$)
- Animal and clinical evaluation

COMPUTER-AIDED LIGAND DESIGN

Aims to reduce number of compounds synthesized and assayed
- Lower costs
- Reduce chemical waste
- Facilitate faster progress

Two main approaches:
1. Receptor/Target-Based
2. Ligand/Drug-Based

SCENARIO 1: RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: Structure-Based Drug Discovery

PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

Docking software
- Search for structure of lowest energy
- Potential function
- Energy as function of structure
- VDW
- Screened Coulombic
- Dihedral

HIV-Protease/KNI-272 complex
STRUCTURE-BASED VIRTUAL SCREENING

- Compound database
- Virtual screening (e.g., computational docking)
  - Candidate ligands
  - Experimental assay
- Ligand optimization
  - Med chem, crystallography, modeling
- Drug candidates

3D structure of target (crystallography, NMR, modeling)

FRAGMENTAL STRUCTURE-BASED SCREENING

- "Fragment" library
- 3D structure of target
- Fragment docking
- Compound design
- Experimental assay and ligand optimization
  - Med chem, crystallography, modeling
- Drug candidates

Multiple non active-site pockets identified

Small organic probe fragment affinities map multiple potential binding sites across the structural ensemble.

Ensemble docking & candidate inhibitor testing

Top hits from ensemble docking against distal pockets were tested for inhibitory effects on basal ERK activity in glioblastoma cell lines.


Proteins and Ligand are Flexible

Protein

\[ \Delta G^\circ \]

Complex
COMMON SIMPLIFICATIONS USED IN PHYSICS-BASED DOCKING

- Quantum effects approximated classically
- Protein often held rigid
- Configurational entropy neglected
- Influence of water treated crudely

Two main approaches:
1. Receptor/Target-Based
2. Ligand/Drug-Based

Scenario 2
Structure of Targeted Protein Unknown: Ligand-Based Drug Discovery

Why Look for Another Ligand if You Already Have Some?

Experimental screening generated some ligands, but they don’t bind tightly.
A company wants to work around another company’s chemical patents.
An high-affinity ligand is toxic, is not well-absorbed, etc.

LIGAND-BASED VIRTUAL SCREENING

CHEMICAL SIMILARITY LIGAND-BASED DRUG-DISCOVERY

Compound Library  \(\rightarrow\) Known Ligands

Molecular similarity
Machine-learning
Etc.

Candidate ligands  \(\rightarrow\) Assay

Activies  \(\rightarrow\) Potent drug candidates

Compounds (available/synthesizable)

Different  \(\rightarrow\) Don’t bother

Similar  \(\rightarrow\) Test experimentally
CHEMICAL FINGERPRINTS

BINARY STRUCTURE KEYS

<table>
<thead>
<tr>
<th>Molecule 1</th>
<th>Molecule 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenyl</td>
<td>Phenyl</td>
</tr>
<tr>
<td>Methyl</td>
<td>Methyl</td>
</tr>
<tr>
<td>Ketone</td>
<td>Ketone</td>
</tr>
<tr>
<td>Carboxylate</td>
<td>Carboxylate</td>
</tr>
<tr>
<td>Amide</td>
<td>Amide</td>
</tr>
<tr>
<td>Aldehyde</td>
<td>Aldehyde</td>
</tr>
<tr>
<td>Chlorine</td>
<td>Chlorine</td>
</tr>
<tr>
<td>Fluorine</td>
<td>Fluorine</td>
</tr>
<tr>
<td>Ethyl</td>
<td>Ethyl</td>
</tr>
<tr>
<td>Naphthyl</td>
<td>Naphthyl</td>
</tr>
</tbody>
</table>

CHEMICAL SIMILARITY FROM FINGERPRINTS

Tanimoto Similarity or Jaccard Index, \( T = \frac{N_I}{N_U} = 0.25 \)

Intersection

Union

Molecule 1

Molecule 2

POTENTIAL DRAWBACKS OF PLAIN CHEMICAL SIMILARITY

May miss good ligands by being overly conservative

May put too much weight on irrelevant details
- Examine ligand shape and common substructures
- Build pharmacophore models
- Statistics and machine learning on chemical descriptors

Maximum Common Substructure

\( N_{\text{common}} = 34 \)

PHARMACOPHORE MODELS

Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore

Bulky hydrophobe

Aromatic

\( 5.0 \pm 0.3 \text{ Å} \)

\( 5.2 \pm 0.4 \text{ Å} \)

\( 2.8 \pm 0.3 \text{ Å} \)

MOLECULAR DESCRIPTORS

More abstract than chemical fingerprints

Physical descriptors
- molecular weight
- charge
- dipole moment
- number of H-bond donors/acceptors
- number of rotatable bonds
- hydrophobicity (log P and clogP)

Topological
- branching index
- measures of linearity vs interconnectedness

Etc. etc.
A High-Dimensional “Chemical Space”
Each compound is at a point in an n-dimensional space.
Compounds with similar properties are near each other.

CAUTIONARY NOTES
• “Everything should be made as simple as it can be but not simpler”
A model is never perfect. A model that is not quantitatively accurate in every respect does not preclude one from establishing results relevant to our understanding of biomolecules as long as the biophysics of the model are properly understood and explored.

• Calibration of the parameters is an ongoing and imperfect process
Questions and hypotheses should always be designed such that they do not depend crucially on the precise numbers used for the various parameters.

• A computational model is rarely universally right or wrong
A model may be accurate in some regards, inaccurate in others. These subtleties can only be uncovered by comparing to all available experimental data.

SUMMARY
• Structural bioinformatics is computer aided structural biology
• Described major motivations, goals and challenges of structural bioinformatics
• Reviewed the fundamentals of protein structure
• Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally