Key concept:

Potential functions describe a system's energy as a function of its structure.

Two main approaches:

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

1. **Physics-Based**
2. **Knowledge-Based**

For **physics** based potentials
energy terms come from physical theory

\[ V(R) = E_{\text{bonded}} + E_{\text{non.bonded}} \]

Sum of **bonded** and **non-bonded** atom-type and position based terms

\[ E_{\text{bonded}} \] is itself a sum of three terms:
\[ V(R) = E_{\text{bonded}} + E_{\text{non.bonded}} \]

\[ E_{\text{bonded}} \text{ is itself a sum of three terms:} \]

\[ E_{\text{bond.stretch}} + E_{\text{bond.angle}} + E_{\text{bond.rotate}} \]

- **Bond Stretch**
  \[ E_{\text{bond.stretch}} \]

- **Bond Angle**
  \[ E_{\text{bond.angle}} \]

- **Bond Rotate**
  \[ E_{\text{bond.rotate}} \]

\[ \sum_{\text{bonds}} K_{bs} (b_i - b_o) \]

\[ \sum_{\text{angles}} K_{ba} (\theta_i - \theta_o) \]

\[ \sum_{\text{dihedrals}} K_{br} [1 - \cos(n_i \phi_i - \phi_o)] \]
\[ V(R) = E_{bonded} + E_{non.bonded} \]

\( E_{non.bonded} \) is a sum of two terms:

\[ E_{van.der.Waals} + E_{electrostatic} \]
Total potential energy

The potential energy can be given as a sum of terms for: Bond stretching, Bond angles, Bond rotations, van der Walls and Electrostatic interactions between atom pairs

\[ V(R) = E_{\text{bond.stretch}} + E_{\text{bond.angle}} + E_{\text{bond.rotate}} + E_{\text{van.der.Waals}} + E_{\text{electrostatic}} \]

Potential energy surface

Now we can calculate the potential energy surface that fully describes the energy of a molecular system as a function of its geometry

\[ E_{\text{electrostatic}} = \sum_{\text{pairs},i,j} \frac{q_i q_j}{e r_{ij}} \]

\[ E_{\text{van.der.Waals}} = \sum_{\text{pairs},i,j} \left[ e_{ij} \left( \frac{r_{oij}}{r_{ij}} \right)^{12} - 2 e_{ij} \left( \frac{r_{oij}}{r_{ij}} \right)^6 \right] \]
**Key concept:**

Now we can calculate the potential energy surface that fully describes the energy of a molecular system as a function of its geometry.

- The forces are the gradients of the energy
  \[ F(x) = -\frac{dV}{dx} \]

**Moving Over The Energy Surface**

- **Energy Minimization**
  Drops into local minimum

- **Molecular Dynamics**
  Uses thermal energy to move smoothly over surface

- **Monte Carlo Moves**
  Random. Accept with probability:
  \[ \exp(-\Delta V/dx) \]

**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 approxs
  - Force fields, quantum
  - Entropy, water effects
- Moore’s law: hardware improving
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
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 $p(r_{O,N})$
3. Compute $E(r_{O,N})$ from $p(r_{O,N})$

ENERGY DETERMINES PROBABILITY (STABILITY)

Basic idea: Use probability as a proxy for energy

Boltzmann:
\[ p(r) \propto e^{-E(r)/RT} \]

Inverse Boltzmann:
\[ E(r) = -RT \ln[p(r)] \]

Computer Aided Drug Discovery

Next Up:
- 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
- Drug discovery & Predicting functional dynamics

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)
THE TRADITIONAL EMPIRICAL PATH TO DRUG DISCOVERY

- Compound library (commercial, in-house, synthetic, natural)
- High throughput screening (HTS)
- Hit confirmation
- Lead compounds (e.g., μM $K_d$)
- 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

HIV Protease/KNI-272 complex

**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

**STRUCTURE-BASED VIRTUAL SCREENING**

Compound database

Virtual screening (e.g., computational docking)

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

Candidate ligands

Ligand optimization

Med chem, crystallography, modeling

Experimental assay

Ligands

Drug candidates

**COMPOUND LIBRARIES**

Commercial (in-house pharma)

Government (NIH)

Academia
COMMON SIMPLIFICATIONS USED IN PHYSICS-BASED DOCKING

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

Do it Yourself!
Hand-on time!

https://bioboot.github.io/bgg213_S19/lectures/#12

You can use the classroom computers or your own laptops. If you are using your laptops then you will need to install MGLTools

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

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

e.g. MAP Kinase Inhibitors

Using knowledge of existing inhibitors to discover more
Why Look for Another Ligand if You Already Have Some?

Experimental screening generated some ligands, but they don’t bind tightly enough

A company wants to work around another company’s chemical patents

An high-affinity ligand is toxic, is not well-absorbed, difficult to synthesize etc.
**CHEMICAL SIMILARITY FROM FINGERPRINTS**

Molecule 1
- phenyl
- methyl
- aldehyde
- carboxylate
- S-S bond
- alcohol

Molecule 2
- methyl
- chloro
- ethyl
- naphthyl
- phenyl

**Tanimoto Similarity (or Jaccard Index), T**

\[ T = \frac{N_I}{N_U} = 0.25 \]

Intersection
- phenyl
- methyl
- aldehyde

Union
- methyl
- chloro
- ethyl
- naphthyl
- phenyl

**Pharmacophore Models**

Φάρμακο (drug) + Φορά (carry)

A 3-point pharmacophore

- Bulky hydrophobe
  - 5.0 ±0.3 Å
- Aromatic
  - 3.2 ±0.4 Å
- +1
  - 2.8 ±0.3 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 a point in an n-dimensional space

Compounds with similar properties are near each other

Apply multivariate statistics and machine learning for descriptor selection. (e.g. partial least squares, PCA, support vector machines, random forest, deep learning etc.)
Proteins and Ligand are Flexible

Protein + Ligand → Complex

ΔG°

Normal Mode Analysis (NMA) is a bioinformatics method to predict the intrinsic dynamics of biomolecules.

Do it Yourself!

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

```r
pdb <- read.pdb("1hel")
modes <- nma(pdb)
m7 <- mktrj(modes, mode=7, file="mode_7.pdb")

library("bio3d.view")
view(m7, col=vec2color(rmsf(m7)))
```

Then you can open the resulting `mode_7.pdb` file in VMD
- Use "TUBE" representation and hit the play button...

Or use the `bio3d.view view()` function

Reference Slides

Molecular Dynamics (MD) and Normal Mode Analysis (NMA) Background and Cautionary Notes

[ Muddy Point Assessment ]
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=ui1ZysMFcKk ]
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)

Nucleic motion described classically

$\frac{d}{dt} \hat{R} = -\nabla_{\hat{R}} E(\hat{R})$

Empirical force field

$E(\hat{R}) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} k_{ij} |\hat{R}_{ij}| - \sum_{i=1}^{n} k_{i} |\hat{R}_{i}|$

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

$\frac{v(t + \Delta t)}{2} = v(t - \frac{\Delta t}{2}) + \frac{F(t)}{m} \Delta t$

$r(t + \Delta t) = r(t) + v(t + \frac{\Delta t}{2}) \Delta t$

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

---

MD Prediction of Functional Motions

"close"

"open"

Yao and Grant, Biophys J. (2013)
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

C. G.

Coarse Grained

1 bead / 1 amino acid
Connected by springs

INFORMING SYSTEMS BIOLOGY?

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

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 structural bioinformatics analysis
- Introduced both physics and knowledge based modeling approaches for describing the structure, energetics and dynamics of proteins computationally
- Introduced both structure and ligand based bioinformatics approaches for drug discovery and design

Ilan Samish et al. Bioinformatics 2015;31:146-150
CAUTIONARY NOTES

• 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 parameters is an ongoing 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.