BGGN 213

Structural Bioinformatics II

Lecture 12

Barry Grant

UC San Diego

http://thegrantlab.org/bggn213

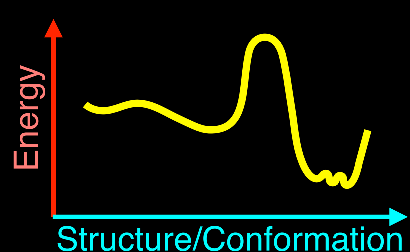
Download MGL Tools: See class website!

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

Key concept:

Potential functions describe a systems 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

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

$$V(R) = E_{bonded} + E_{non.bonded}$$

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

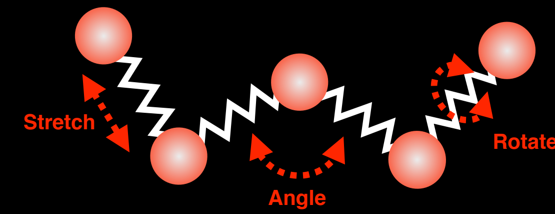
$$V(R) = \boxed{E_{bonded}} + E_{non.bonded}$$

E_{bonded} is itself a sum of three terms:

$$V(R) = \boxed{E_{bonded}} + E_{non.bonded}$$

E_{bonded} is itself a sum of three terms:

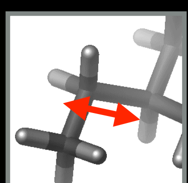
$$\boxed{E_{bond.stretch} + E_{bond.angle} + E_{bond.rotate}}$$



$$V(R) = \boxed{E_{bonded}} + E_{non.bonded}$$

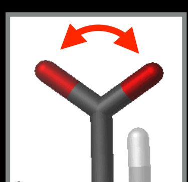
E_{bonded} is itself a sum of three terms:

$$\boxed{E_{bond.stretch} + E_{bond.angle} + E_{bond.rotate}}$$



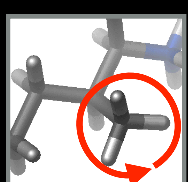
Bond Stretch

$E_{bond.stretch}$



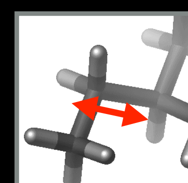
Bond Angle

$E_{bond.angle}$



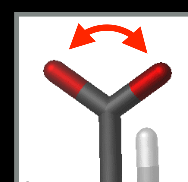
Bond Rotate

$E_{bond.rotate}$



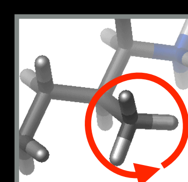
Bond Stretch

$$\sum_{bonds} K_i^{bs}(b_i - b_o)$$



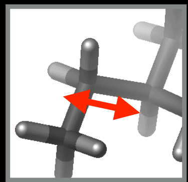
Bond Angle

$$\sum_{angles} K_i^{ba}(\theta_i - \theta_o)$$



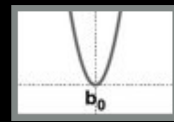
Bond Rotate

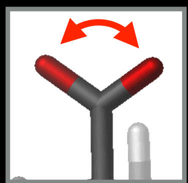
$$\sum_{dihedrals} K_i^{br}[1 - cos(n_i\phi_i - \phi_o)$$



Bond Stretch

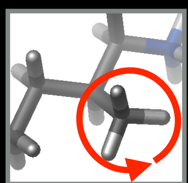
$$\sum_{bonds} K_i^{bs}(b_i - b_o)$$





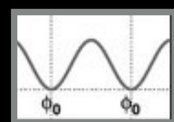
Bond Angle

$$\sum_{angles} K_i^{ba}(\theta_i - \theta_o)$$



Bond Rotate

$$\sum_{dihedrals} K_i^{br}[1 - cos(n_i\phi_i - \phi_o)]$$



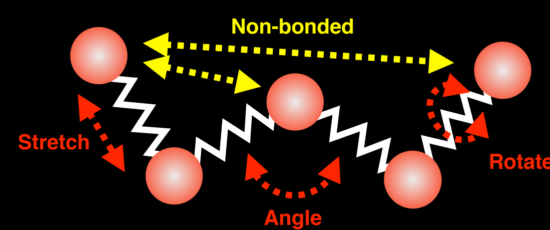
$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

$E_{non.bonded}$ is a sum of two terms:

$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

$E_{non.bonded}$ is a sum of two terms:

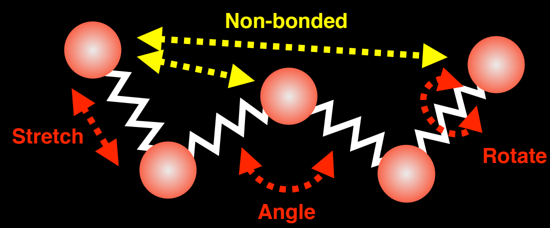
$$E_{van.der.Waals} + E_{electrostatic}$$



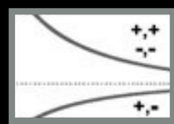
$$V(R) = E_{bonded} + \boxed{E_{non.bonded}}$$

$E_{non.bonded}$ is a sum of two terms:

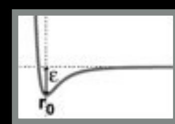
$$E_{van.der.Waals} + E_{electrostatic}$$



$$E_{electrostatic} = \sum_{pairs.i.j} \frac{q_i qj}{\epsilon r_{ij}}$$



$$E_{van.der.Waals} = \sum_{pairs.i.j} \left[\epsilon_{ij} (\frac{r_{o.ij}}{r_{ij}})^{12} - 2\epsilon_{ij} (\frac{r_{o.ij}}{r_{ij}})^6 \right]$$



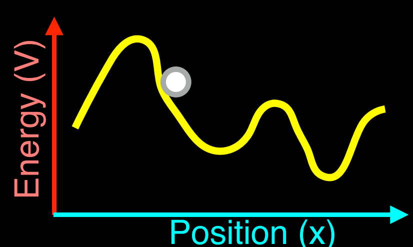
Total potential energy

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

$$\begin{aligned} V(R) = & \left. \begin{array}{l} E_{bond.stretch} \\ +E_{bond.angle} \\ +E_{bond.rotate} \end{array} \right\} E_{bonded} \\ & \left. \begin{array}{l} +E_{van.der.Waals} \\ +E_{electrostatic} \end{array} \right\} E_{non.bonded} \end{aligned}$$

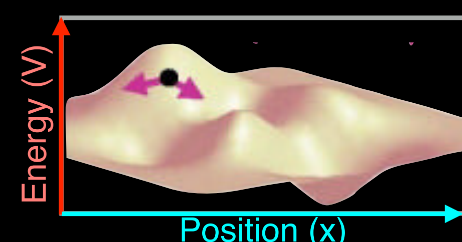
Potential energy surface

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



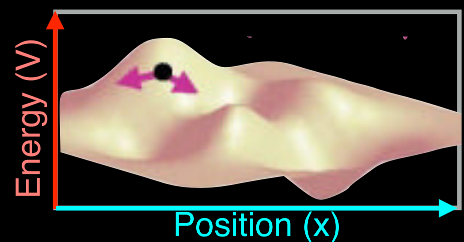
Potential energy surface

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



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



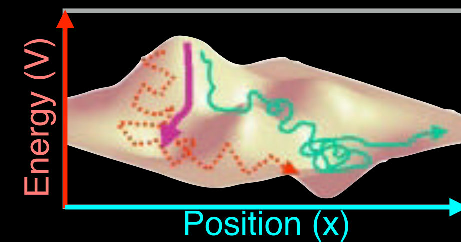
- The **forces** are the gradients of the energy

$$F(x) = -dV/dx$$

Moving Over The Energy Surface

- **Energy Minimization** drops into local minimum
- **Molecular Dynamics** uses thermal energy to move smoothly over surface
- **Monte Carlo Moves** are random. Accept with probability:

$$exp(-\Delta V/dx)$$



PHYSICS-ORIENTED APPROACHES

Weaknesses

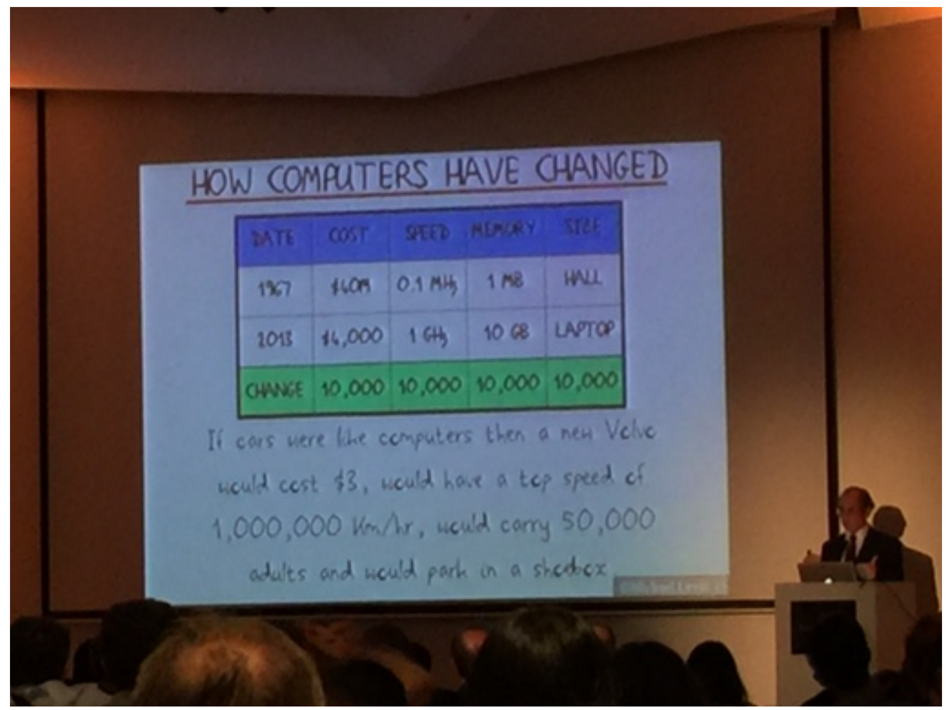
Fully physical detail becomes computationally intractable
Approximations are unavoidable
(Quantum effects approximated classically, water may be treated crudely)
Parameterization still required

Strengths

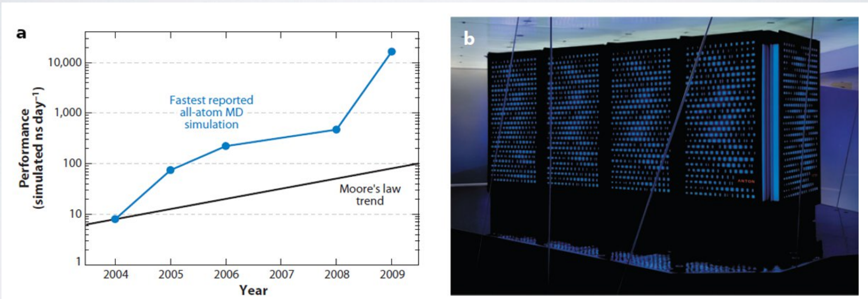
Interpretable, provides guides to design
Broadly applicable, in principle at least
Clear pathways to improving accuracy

Status

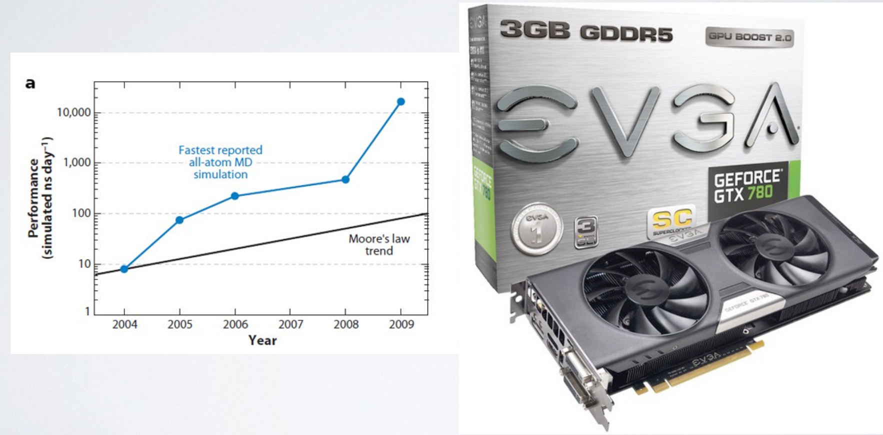
Useful, widely adopted but far from perfect
Multiple groups working on fewer, better approxs
Force fields, quantum
entropy, water effects
Moore's law: hardware improving



SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



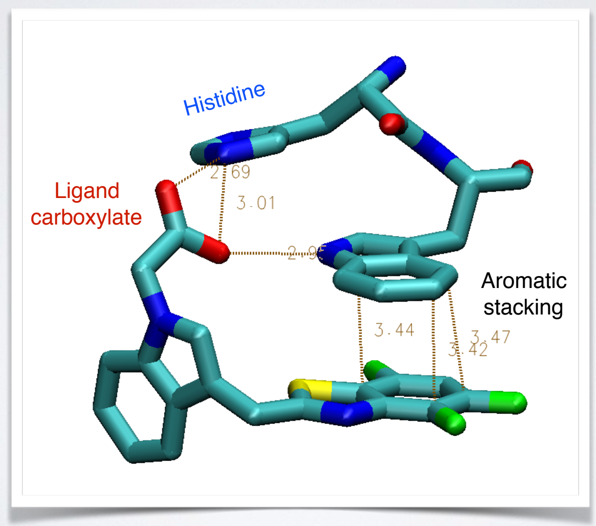
POTENTIAL FUNCTIONS DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION OF ITS **STRUCTURE**

Two main approaches:

(1). Physics-Based

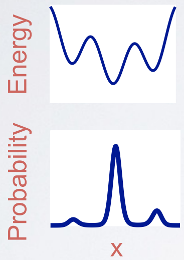
(2). Knowledge-Based

KNOWLEDGE-BASED DOCKING POTENTIALS



ENERGY DETERMINES **PROBABILITY** (STABILITY)

Basic idea: Use probability as a proxy for energy



Boltzmann:

$$p(r) \propto e^{-E(r)/RT}$$

Inverse Boltzmann:

$$E(r) = -RT \ln\left[p(r)\right]$$

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 $p(r_{O\text{-}N})$
3. Compute $E(r_{O\text{-}N})$ from $p(r_{O\text{-}N})$

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)

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

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

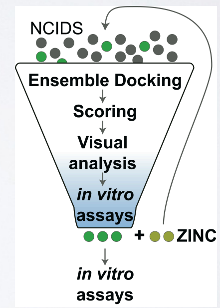
COMPUTER-AIDED LIGAND DESIGN

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

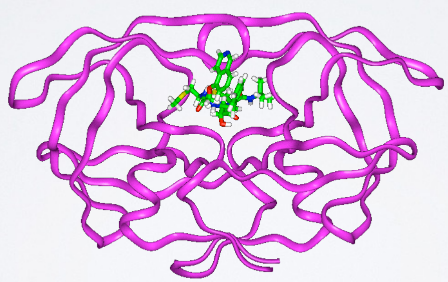
Two main approaches:

(1). Receptor/Target-Based

(2). Ligand/Drug-Based

SCENARIO I:
RECEPTOR-BASED DRUG DISCOVERY

Structure of Targeted Protein Known: Structure-Based Drug Discovery

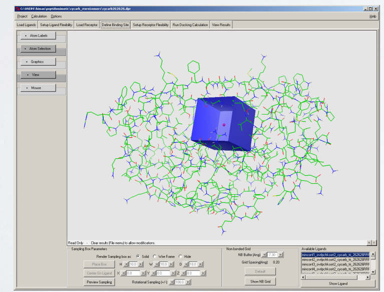


HIV Protease/KNI-272 complex

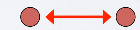
PROTEIN-LIGAND DOCKING

Structure-Based Ligand Design

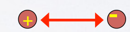
Docking software
Search for structure of lowest energy



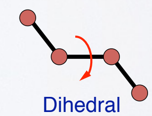
Potential function
Energy as function of structure



VDW



Screened Coulombic



Dihedral

STRUCTURE-BASED VIRTUAL SCREENING

Compound database → Virtual screening

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

Virtual screening (e.g., computational docking)
↓
Candidate ligands
↓
Experimental assay
↓
Ligands → Drug candidates

Ligands → Ligand optimization
Med chem, crystallography, modeling → Candidate ligands

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/bggn213_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

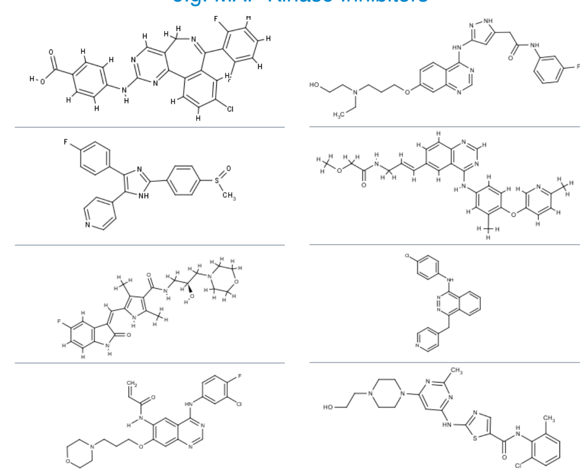
(2). Ligand/Drug-Based

Scenario 2

Structure of Targeted Protein Unknown:

Ligand-Based Drug Discovery

e.g. MAP Kinase Inhibitors



Using knowledge of existing inhibitors to discover more

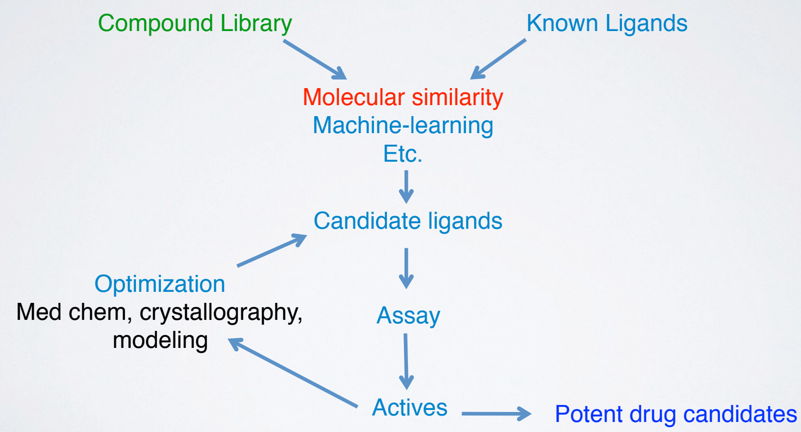
Why Look for Another Ligand if You Already Have Some?

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

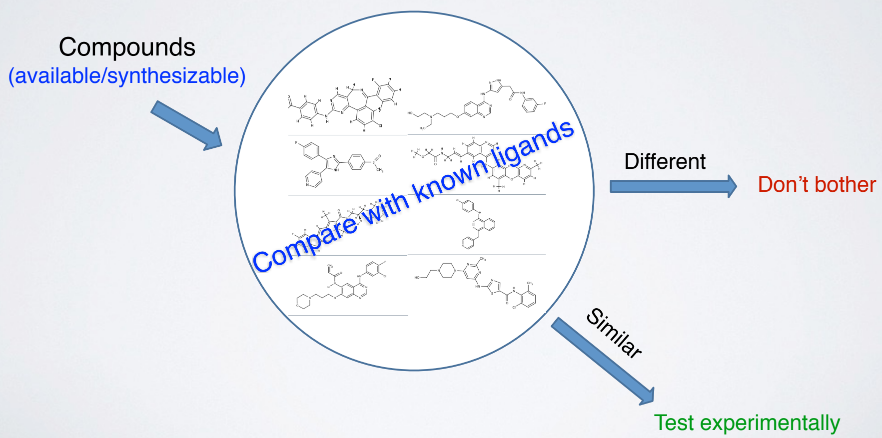
A company wants to work around another company's chemical patents

An high-affinity ligand is toxic, is not well-absorbed, difficult to synthesize etc.

LIGAND-BASED VIRTUAL SCREENING

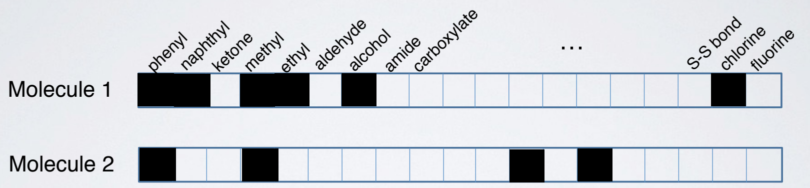


CHEMICAL SIMILARITY LIGAND-BASED DRUG-DISCOVERY



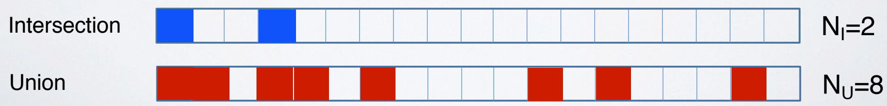
CHEMICAL FINGERPRINTS BINARY STRUCTURE KEYS

CHEMICAL SIMILARITY FROM FINGERPRINTS



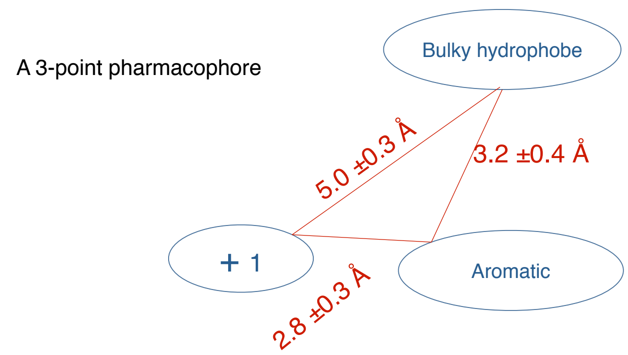
Tanimoto Similarity (or Jaccard Index), T

$$T \equiv \frac{N_I}{N_U} = 0.25$$



Pharmacophore Models

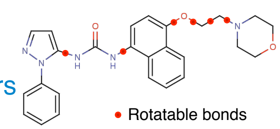
Φάρμακο (drug) + Φορά (carry)



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)



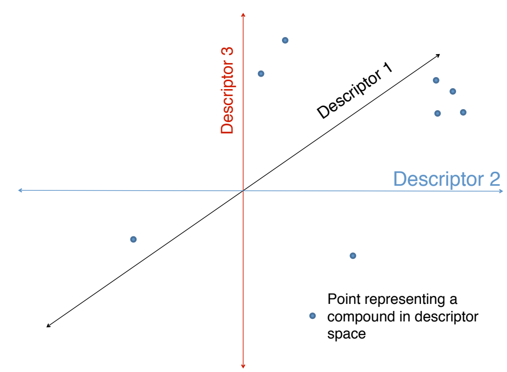
Topological
- branching index
- measures of linearity vs interconnectedness

Etc. etc.

A High-Dimensional "Chemical Space"

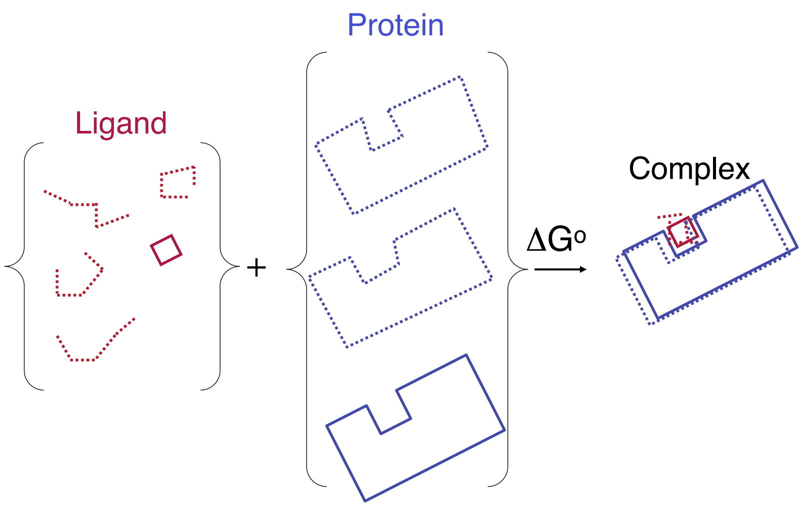
Each compound is a point in an n-dimensional space

Compounds with similar properties are near each other



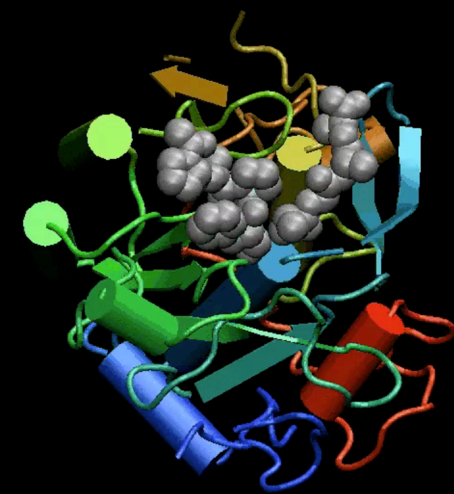
Apply multivariate statistics and machine learning for descriptor-selection. (e.g. partial least squares, PCA, support vector machines, random forest, deep learning etc.)

Proteins and Ligand are Flexible



Do it Yourself!

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



https://bioboot.github.io/bggn213_S19/lectures/#12

NMA in Bio3D

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

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

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

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

Reference Slides

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

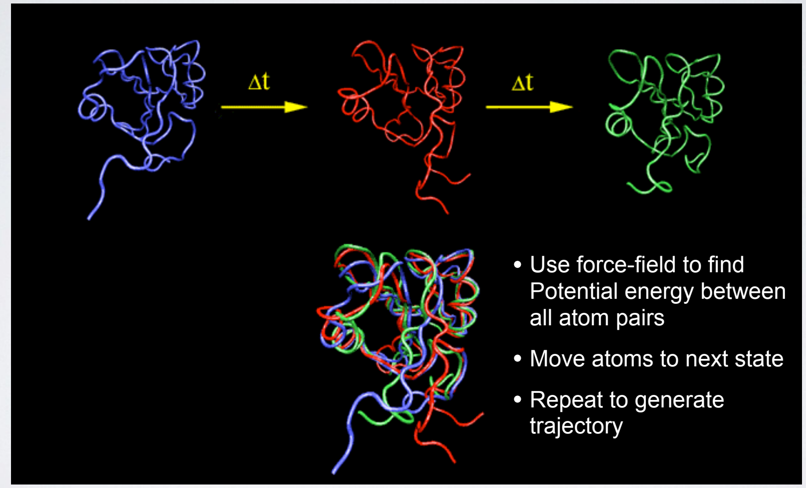
[Muddy Point Assessment]

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.

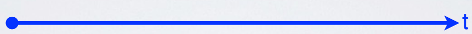
- Thus knowledge of dynamics can provide a deeper understanding of the mapping of structure to function
  - Molecular dynamics (MD) and normal mode analysis (NMA) are two major methods for predicting and characterizing molecular motions and their properties

MOLECULAR DYNAMICS SIMULATION

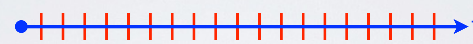


McCammon, Gelin & Karplus, *Nature* (1977)
[See: https://www.youtube.com/watch?v=ui1ZysMFcKk]

- Divide time into discrete (~1fs) time steps (Δt)
  (for integrating equations of motion, see below)

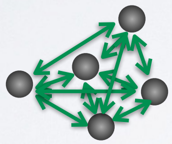


- Divide time into discrete (~1fs) time steps (Δt)
  (for integrating equations of motion, see below)



- Divide time into discrete (~1fs) time steps (Δt)
(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

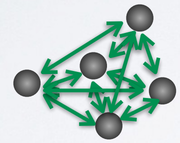
$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical force field

$$E(\vec{R}) = \sum_{\text{bonded}} E_i(\vec{R}) + \sum_{\text{non-bonded}} E_i(\vec{R})$$

- Divide time into discrete (~1fs) time steps (Δt)
(for integrating equations of motion, see below)

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



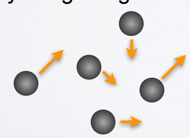
Nucleic motion described classically

$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical force field

$$E(\vec{R}) = \sum_{\text{bonded}} E_i(\vec{R}) + \sum_{\text{non-bonded}} E_i(\vec{R})$$

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

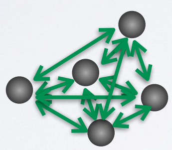


$$\begin{aligned} \boxed{\boldsymbol{v}(t+\frac{\Delta t}{2})} &= \boldsymbol{v}(t-\frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t \\ \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \boxed{\boldsymbol{v}(t+\frac{\Delta t}{2})}\Delta t \end{aligned}$$

BASIC ANATOMY OF A MD SIMULATION

- Divide time into discrete (~1fs) time steps (Δt)
(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

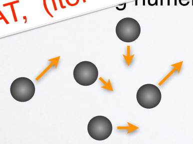
$$m_i \frac{d^2}{dt^2}\vec{R}_i = -\vec{\nabla}_i E(\vec{R})$$

Empirical force f

$$E(\vec{R}) = \sum \ldots E_i(\vec{R})$$

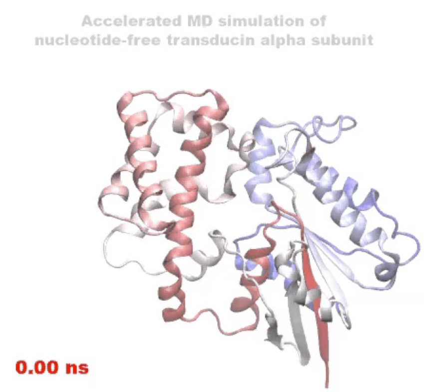
- Use the fo... ...ate velocities and move atoms to new positions
(b... ...g numerically via the “leapfrog” scheme)

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



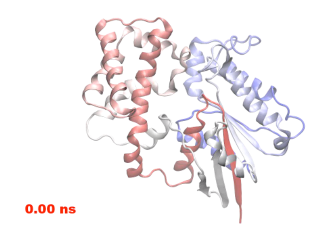
$$\begin{aligned} \boldsymbol{v}(t+\frac{\Delta t}{2}) &= \boldsymbol{v}(t-\frac{\Delta t}{2}) + \frac{\boldsymbol{F}(t)}{m}\Delta t \\ \boldsymbol{r}(t+\Delta t) &= \boldsymbol{r}(t) + \boldsymbol{v}(t+\frac{\Delta t}{2})\Delta t \end{aligned}$$

MD Prediction of Functional Motions

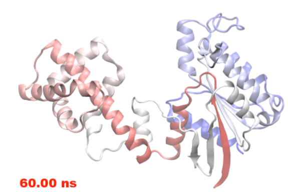


Yao and Grant, Biophys J. (2013)

“close”

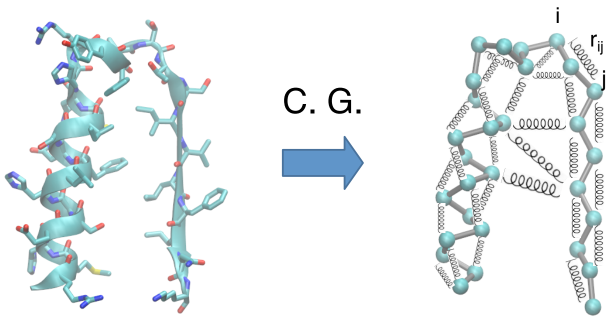


“open”



COARSE GRAINING: **NORMAL MODE ANALYSIS** (NMA)

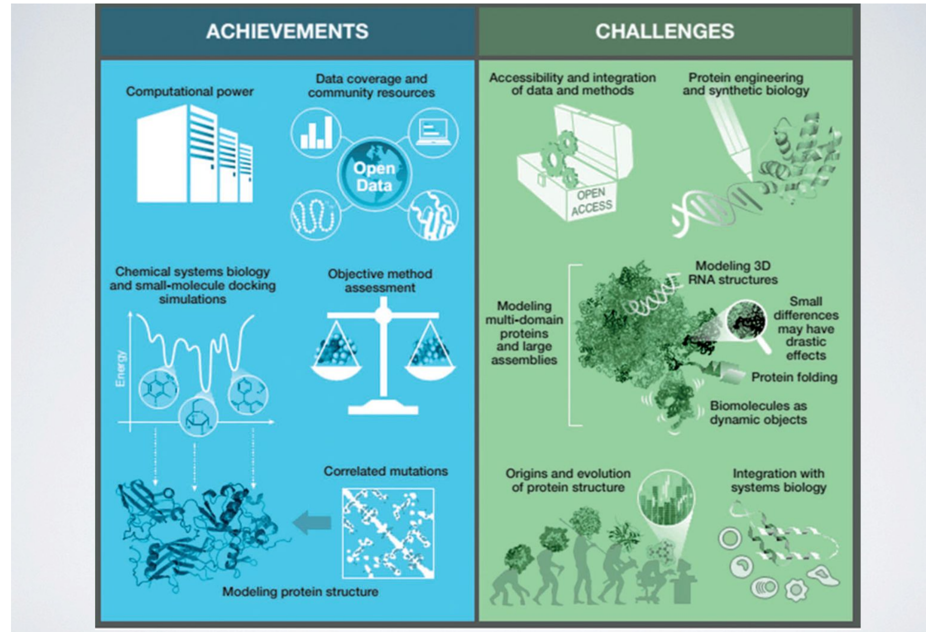
- MD is still time-consuming for large systems
- Elastic network model NMA (ENM-NMA) is an example of a lower resolution approach that finishes in seconds even for large systems.



Atomistic

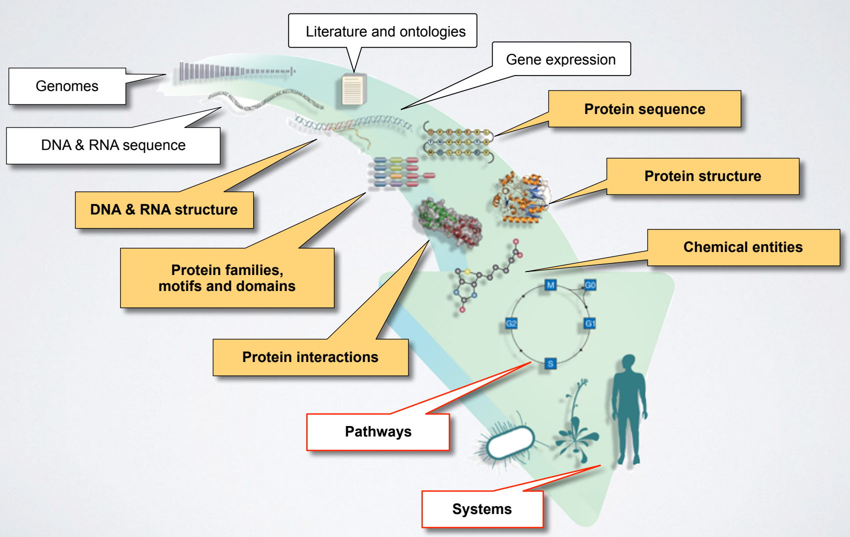
Coarse Grained

- 1 bead / 1 amino acid
- Connected by springs



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

INFORMING SYSTEMS BIOLOGY?



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

[Muddy Point Assessment]

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.