

“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!

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

Bioinformatics is computer aided biology!

Goal: Data to Knowledge

So what is **structural bioinformatics**?

So what is **structural bioinformatics**?

... **computer aided structural biology!**

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

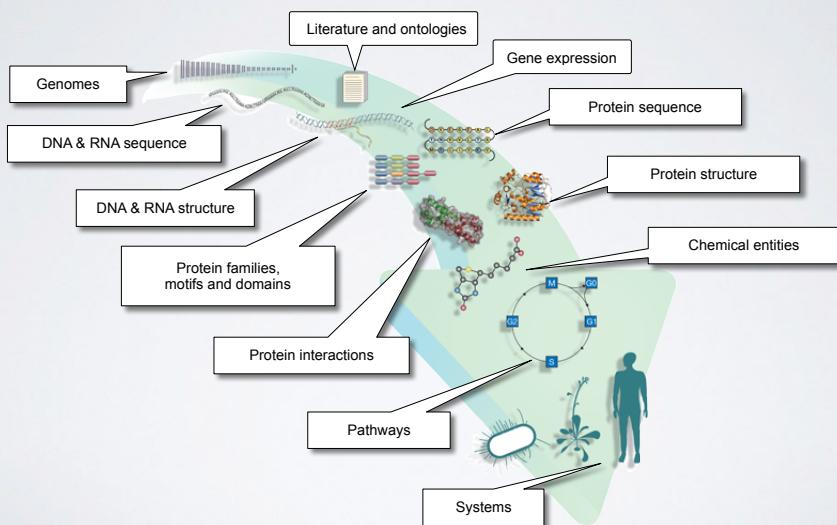
Why should we care?

Why should we care?

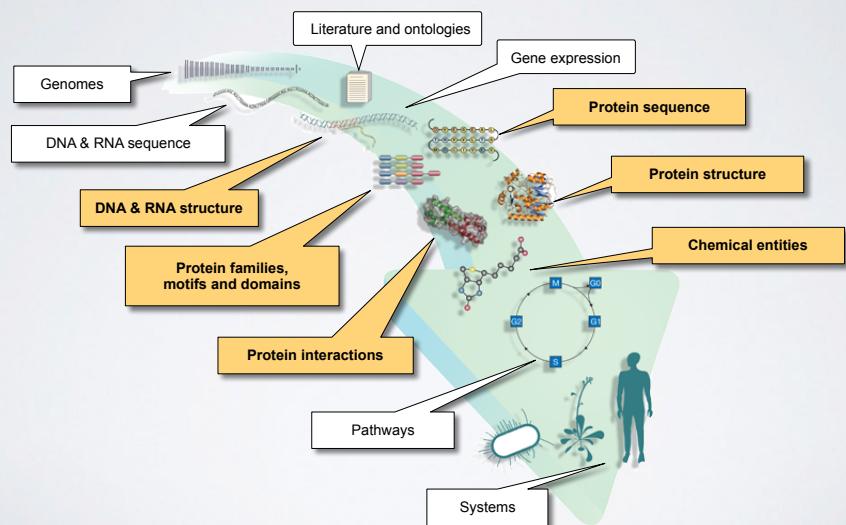
Because biomolecules are “nature’s robots”

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

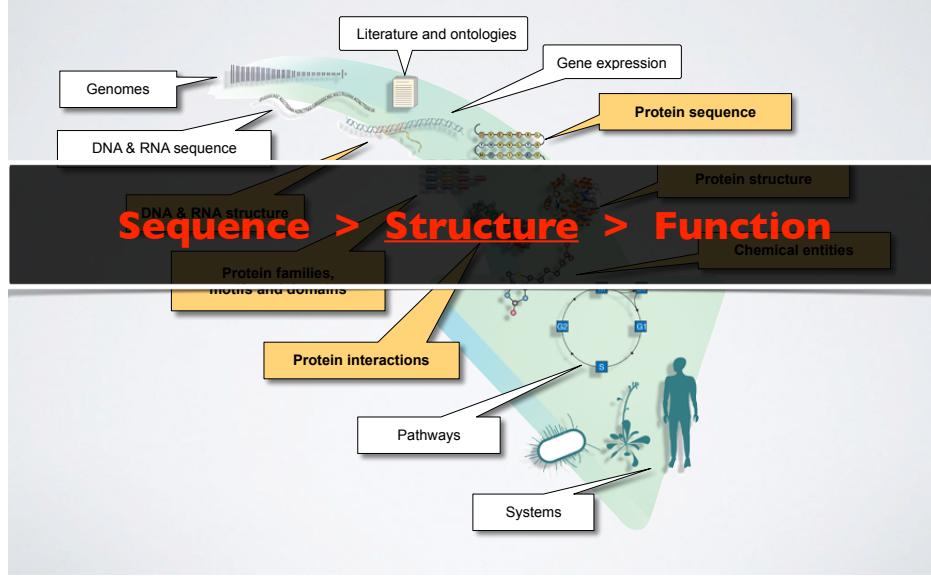
BIOINFORMATICS DATA



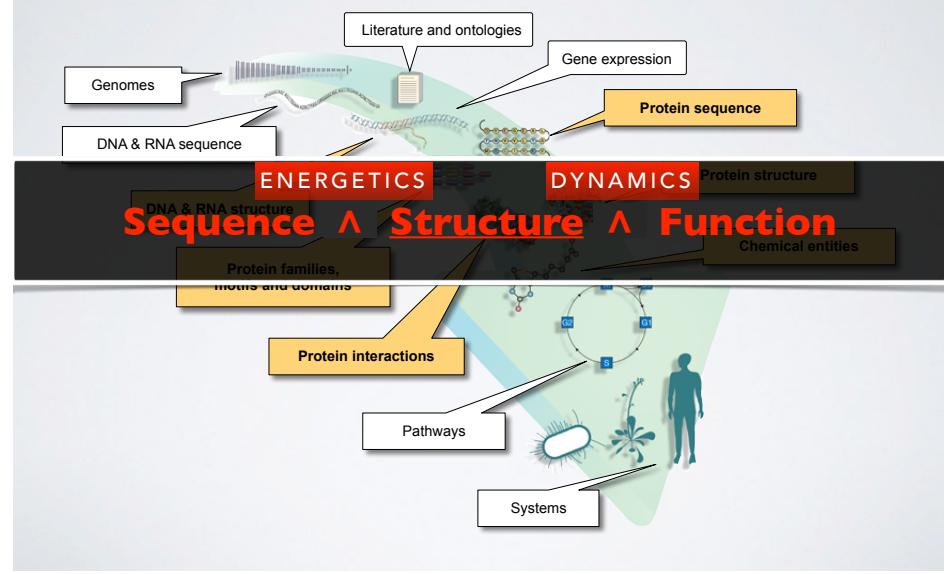
STRUCTURAL DATA IS CENTRAL

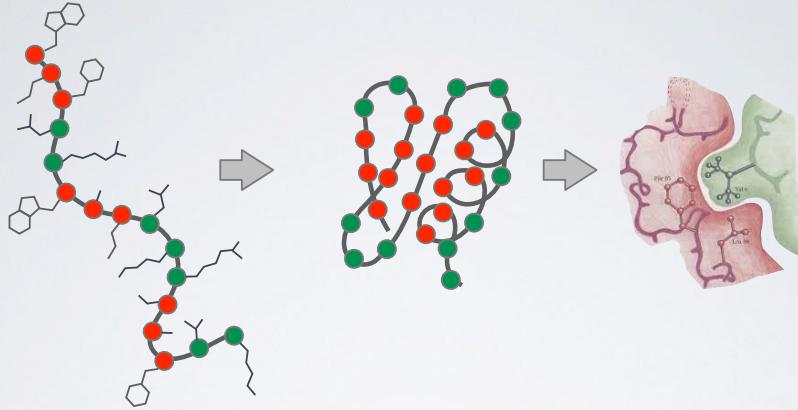


STRUCTURAL DATA IS CENTRAL



STRUCTURAL DATA IS CENTRAL





Sequence	Structure	Function
<ul style="list-style-type: none"> Unfolded chain of amino acid chain Highly mobile Inactive 	<ul style="list-style-type: none"> Ordered in a precise 3D arrangement Stable but dynamic 	<ul style="list-style-type: none"> Active in specific "conformations" Specific associations & precise reactions

In daily life, we use machines with functional *structure* and *moving parts*



Genomics is a great start

Track Bike – DL 175

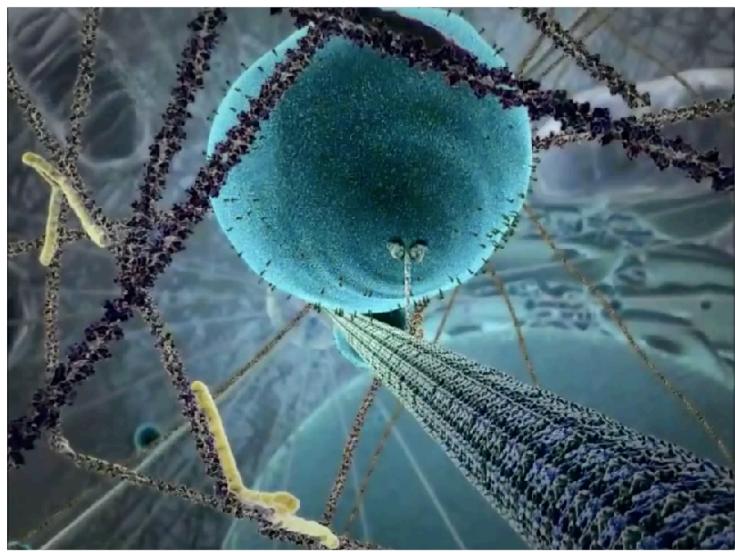
REF. NO.	IBM NO.	DESCRIPTION
1	156011	Track Frame 21", 22", 23", 24", Team Red
2	157040	Fork for 21" Frame
2	157039	Fork for 22" Frame
2	157038	Fork for 23" Frame
2	157037	Fork for 24" Frame
3	191202	Handlebar TTT Competition Track Alloy 15/16" Headset Stem, TTT, Specify extension
4	145978	Examiner Bolt
5	191278	Clamp Bolt
6	191272	Headset Complete 1 x 24 BSC
7	145841	Ball Bearings
8	145842	175 Raleigh Pistard Seta Tubular Prestavalue 27"
9	190420	Rim, 27" AVA Competition (36H) Alloy Prestavalue
10	190233	Hub, Large Flange Campagnolo Pista Track Alloy (pairs)
11	145973	Sprocket, 11 5/8"
12	145974	Sleeve
13	145937	Ball Bearings
14	145636	Bottom Bracket Axle
15	145170	Cone for Sleeve
16	145836	L.H. Adjustable Cup
17	146473	Lockring
18	145833	Straps or Toe Clips
19	145934	Fixed Belt
20	145935	Fixing Washer
21	145822	Dustcap
22	145823	R.H. and L.H. Crankset with Chainwheel
23	146472	Fixed Cup
24	146473	Toe Clips, Christophe, Chrome (Medium)
25	145235	Pedals, Extra Light, Pairs
26	145684	Chain
27	123021	Chainring
28	145980	Seat Post
29	167002	Seat Post Bolt and Nut
30	145933	Saddle, Brooks
31	145933	Track Sprocket, Specify 12, 13, 14, 15, or 16 T.

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

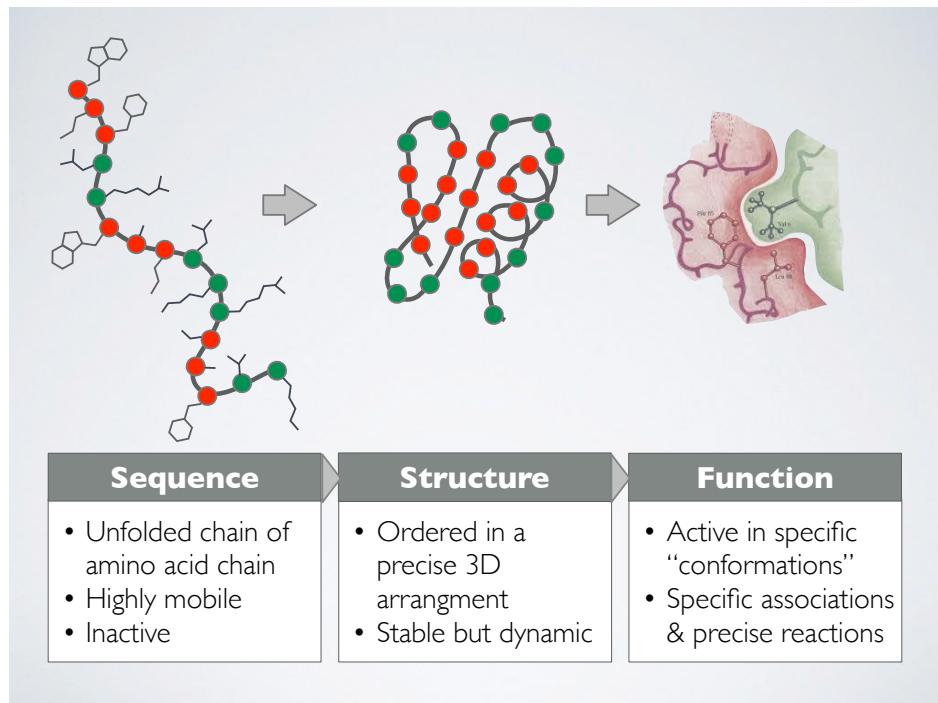
... but not the end



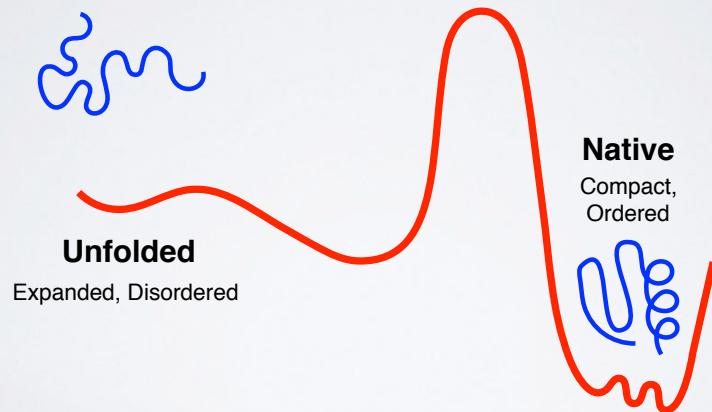
- We want the full spatiotemporal picture, and an ability to control it
- Broad applications, including drug design, medical diagnostics, chemical manufacturing, and energy



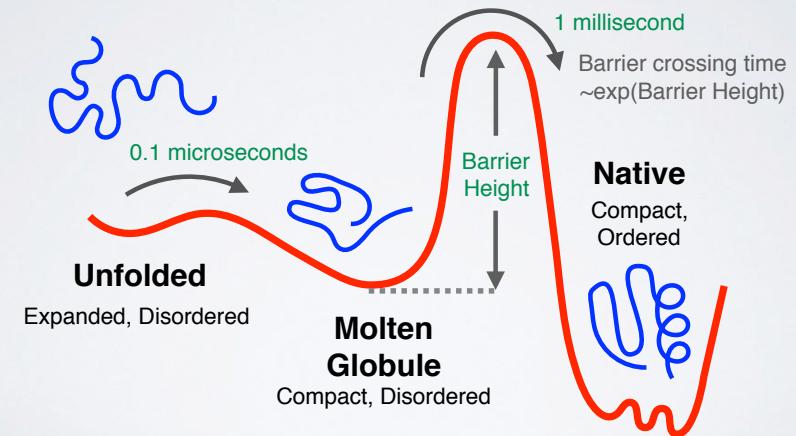
Extracted from The Inner Life of a Cell by Cellular Visions and Harvard
[YouTube link: <https://www.youtube.com/watch?v=y-uuk4Pr2i8>]



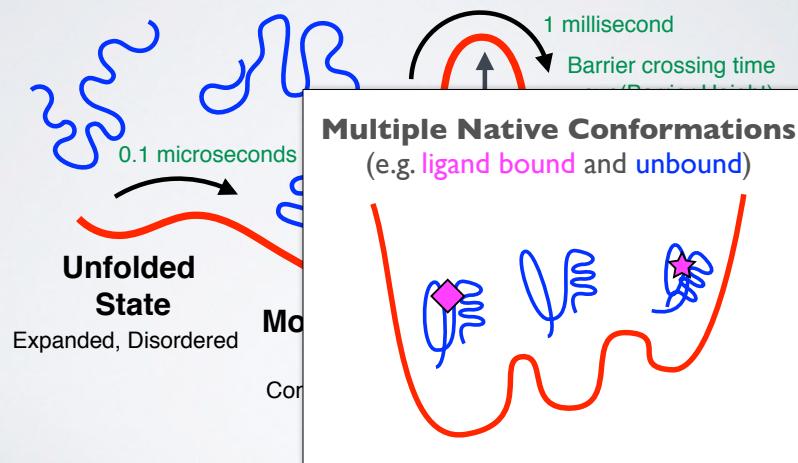
KEY CONCEPT: ENERGY LANDSCAPE



KEY CONCEPT: ENERGY LANDSCAPE



KEY CONCEPT: ENERGY LANDSCAPE



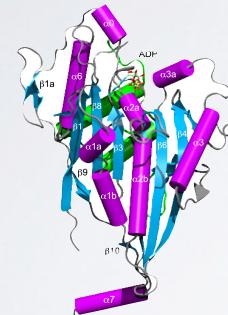
OUTLINE:

- ▶ 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

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)

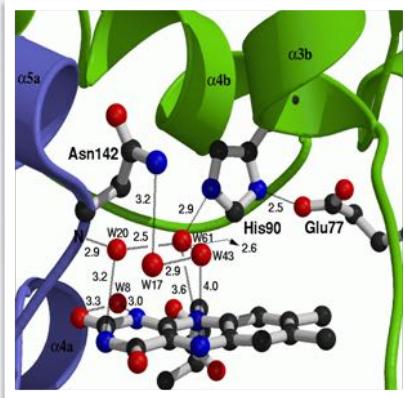


Small Molecules
(CCDB)

Motivation 1:

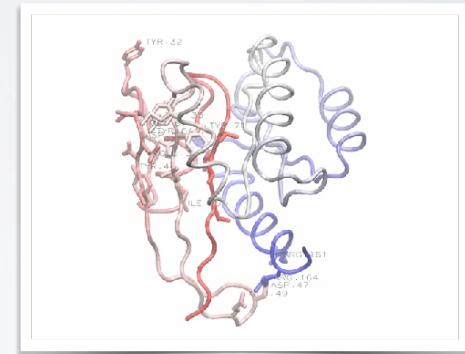
Detailed understanding of molecular interactions

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.

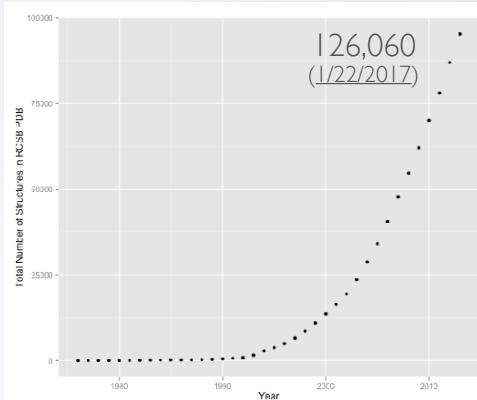


Grant et al. PLoS Comp. Biol. (2010)

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 2:

Lots of structural data is becoming available

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

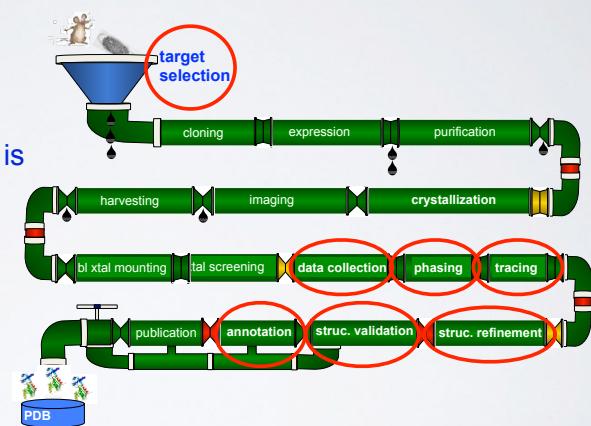
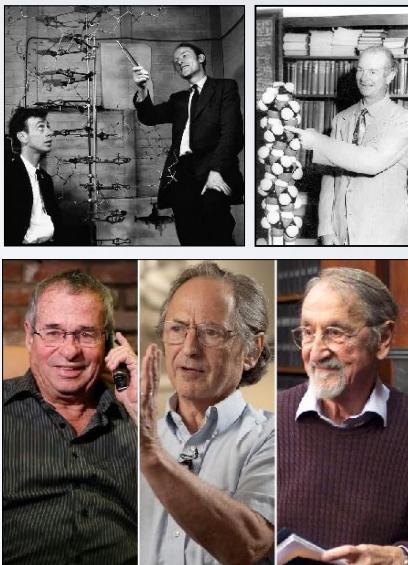
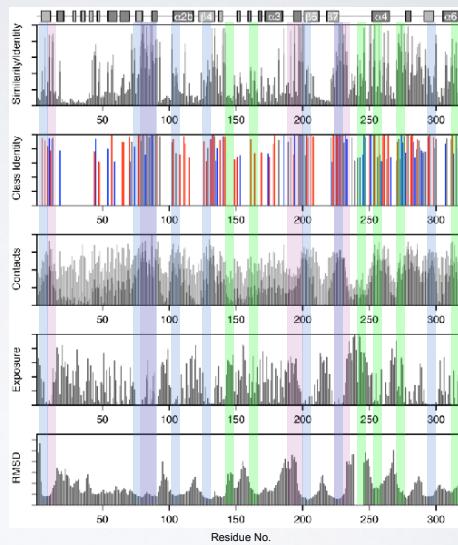


Image Credit: "Structure determination assembly line" Adam Godzik

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



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



Grant et al. JMB. (2007)

SUMMARY OF KEY **MOTIVATIONS**

Sequence > Structure > Function

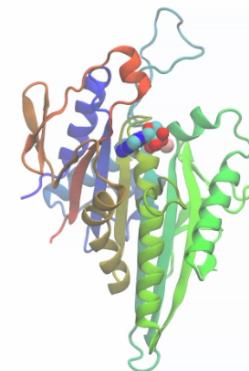
- Structure determines function, so understanding structure helps our understanding of function

Structure is more conserved than sequence

- Structure allows identification of more distant evolutionary relationships

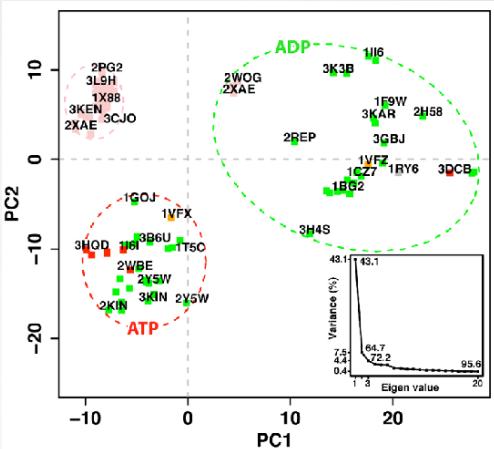
Structure is encoded in sequence

- Understanding the determinants of structure allows design and manipulation of proteins for industrial and medical advantage



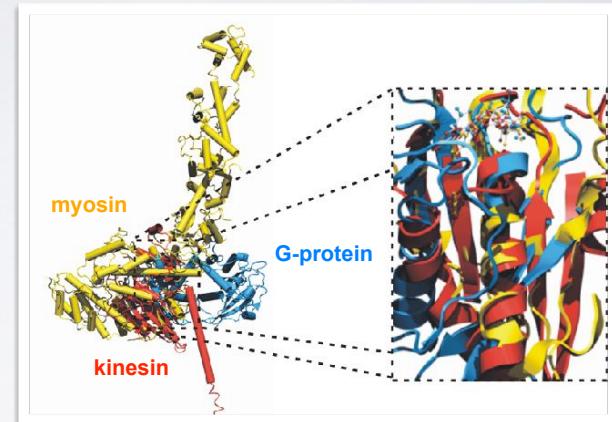
Scarabelli and Grant. PLoS. Comp. Biol. (2013)

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



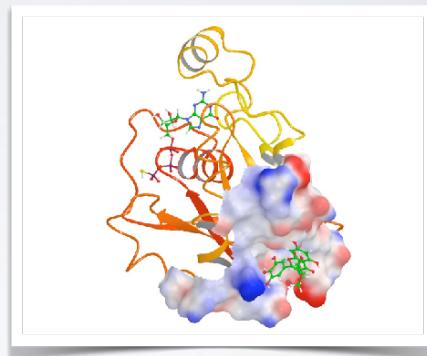
Scarabelli and Grant. PLoS. Comp. Biol. (2013)

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



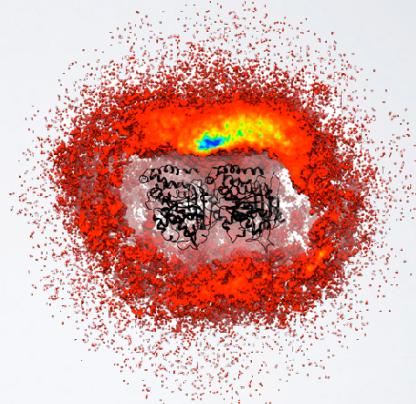
Grant et al. unpublished

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



Grant et al. PLoS One (2011, 2012)

- Goals:
- Analysis
 - Visualization
 - 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

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

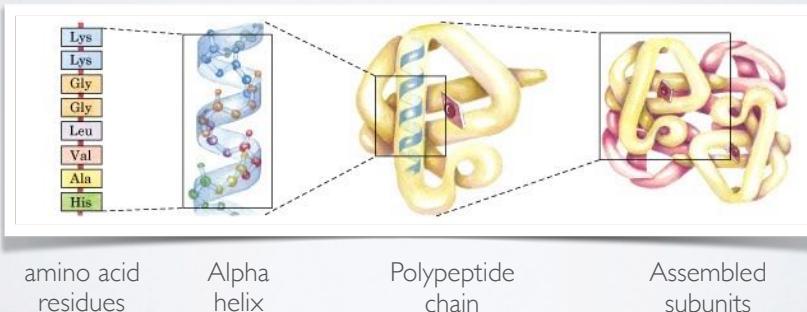


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

RECAP: AMINO ACID NOMENCLATURE

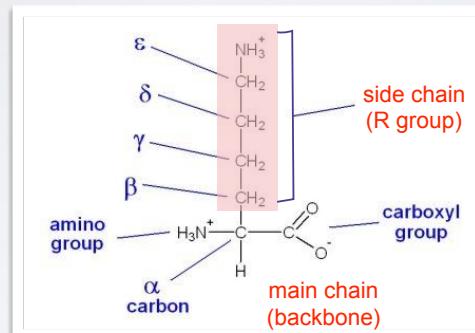


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

AMINO ACIDS CAN BE GROUPED BY THE PHYSIOCHEMICAL PROPERTIES

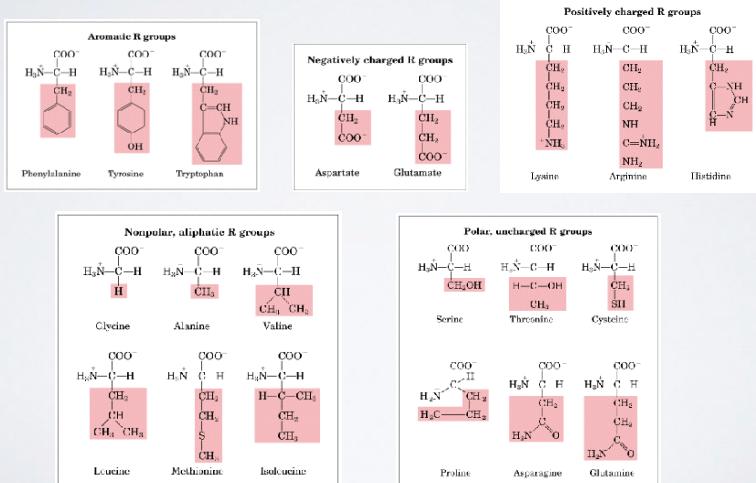


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

AMINO ACIDS POLYMERIZE THROUGH PEPTIDE BOND FORMATION

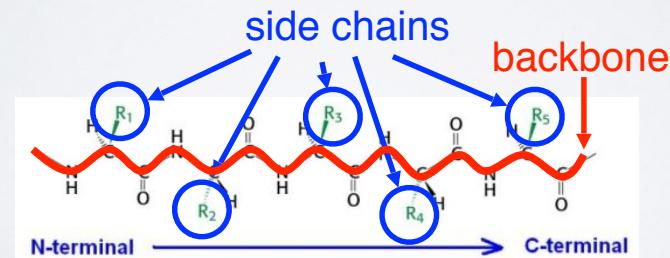
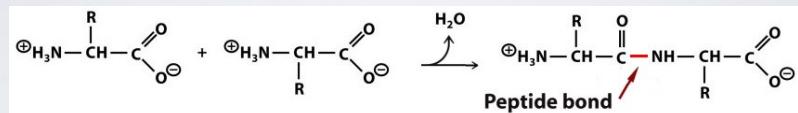


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

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

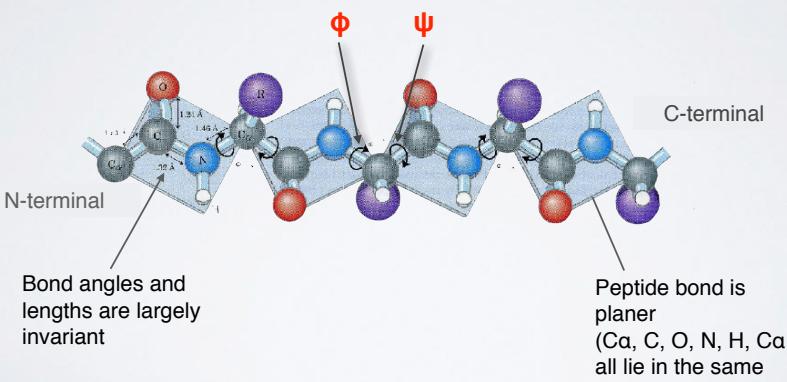
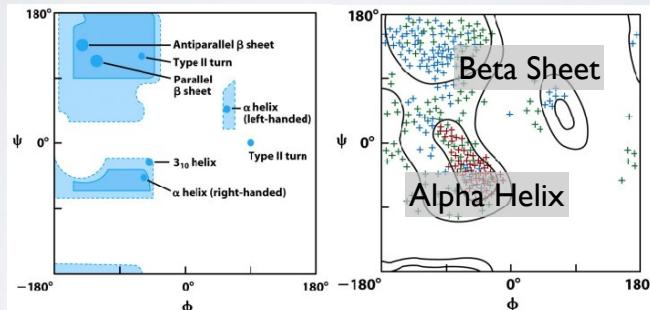


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

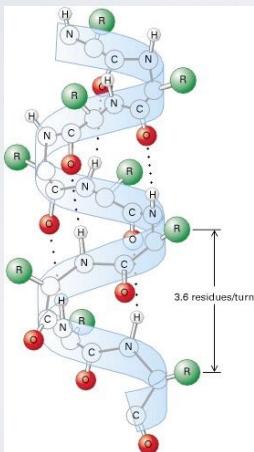
PHI vs PSI PLOTS ARE KNOWN AS RAMACHANDRAN DIAGRAMS



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

Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

MAJOR SECONDARY STRUCTURE TYPES ALPHA HELIX & BETA SHEET

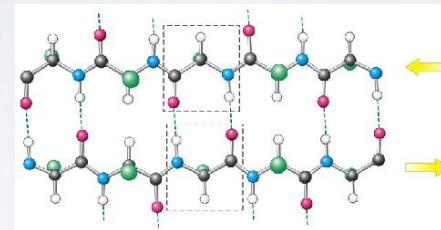


α -helix

- 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

Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

MAJOR SECONDARY STRUCTURE TYPES ALPHA HELIX & **BETA SHEET**

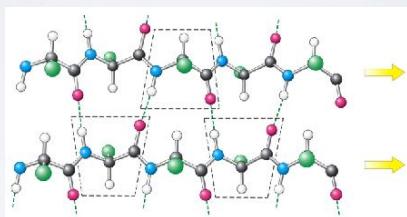


In antiparallel β -sheets

- 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

Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

MAJOR SECONDARY STRUCTURE TYPES ALPHA HELIX & **BETA SHEET**

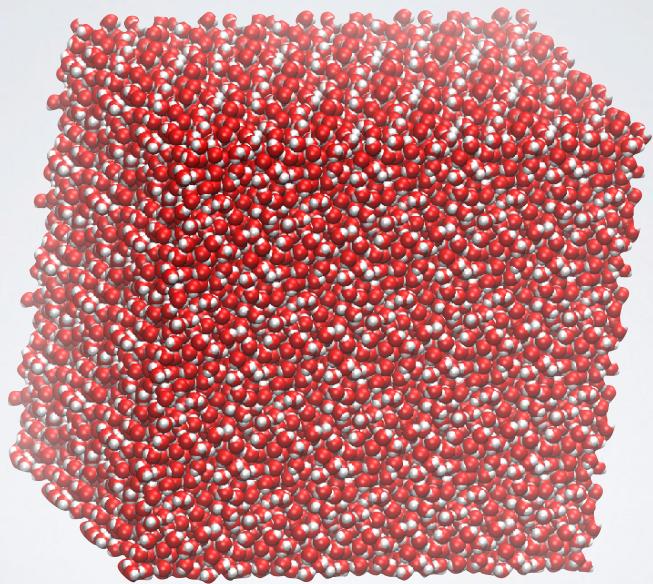


In parallel β -sheets

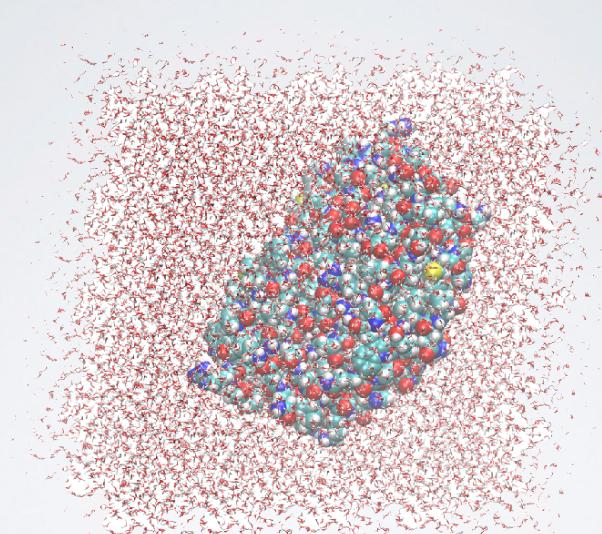
- Adjacent β -strands run in same direction
- Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
- The side chains (in green) are above and below the sheet

Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

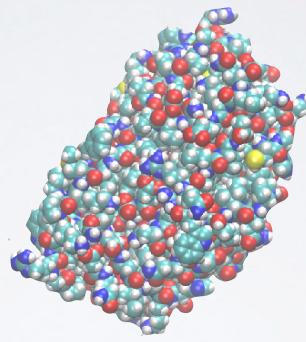
What Does a Protein Look like?



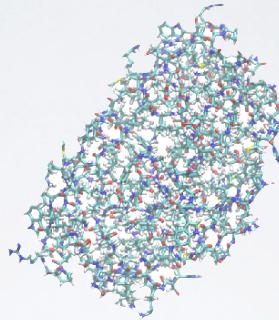
- Proteins are stable (and hidden) in water



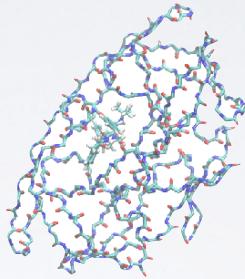
- Proteins closely interact with water



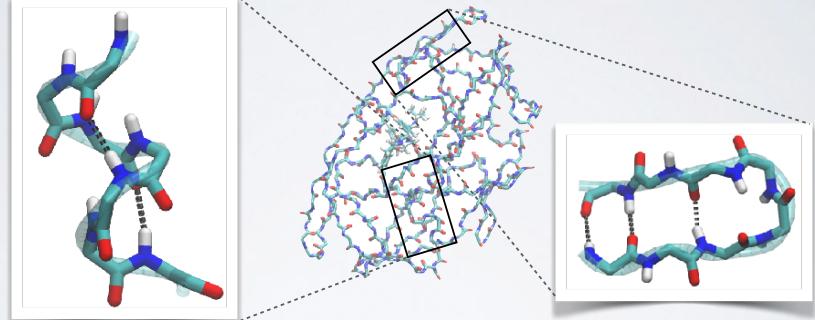
- Proteins are close packed solid but flexible objects (globular)



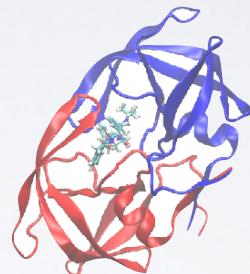
- Due to their large size and complexity it is often hard to see what's important in the structure



- Backbone or main-chain representation can help trace chain topology

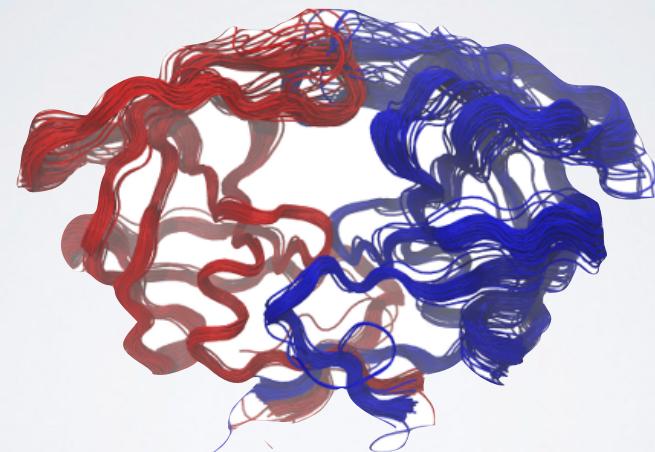


- Backbone or main-chain representation can help trace chain topology & reveal secondary structure



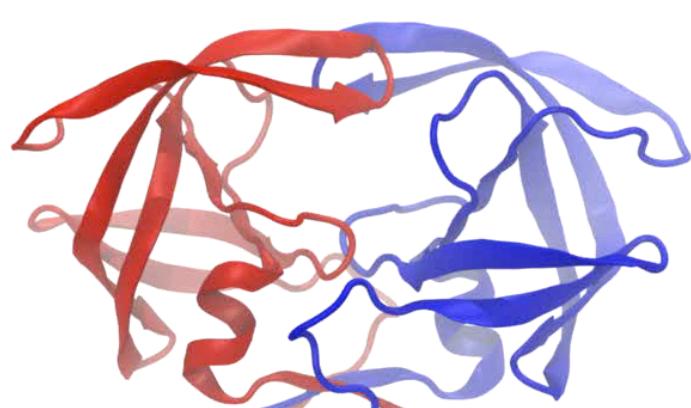
- Simplified secondary structure representations are commonly used to communicate structural details
- Now we can clearly see 2^o, 3^o and 4^o structure
- Coiled chain of connected secondary structures

DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY



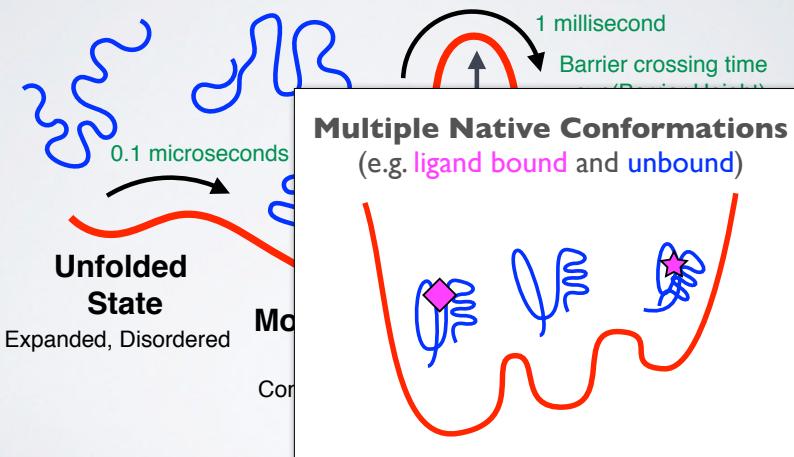
Superposition of all 482 structures in RCSB PDB (23/09/2015)

DISPLACEMENTS REFLECT INTRINSIC FLEXIBILITY



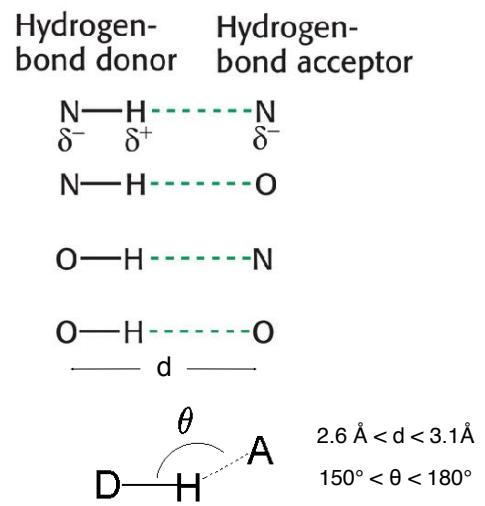
Principal component analysis (PCA) of experimental structures

KEY CONCEPT: ENERGY LANDSCAPE



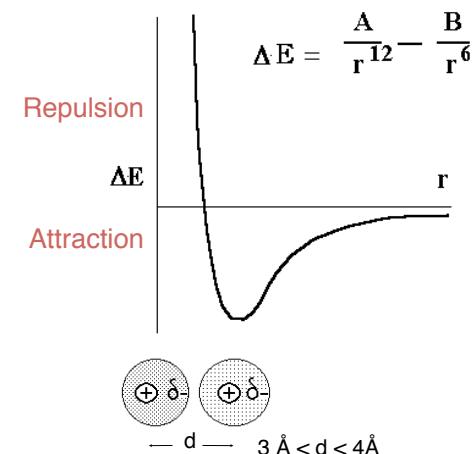
Key forces affecting structure:

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



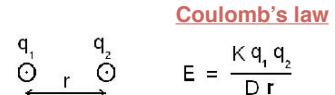
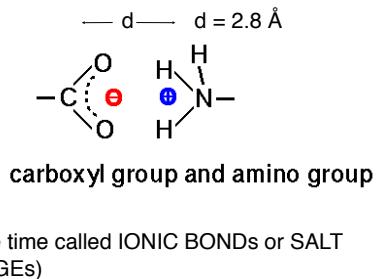
Key forces affecting structure:

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



Key forces affecting structure:

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

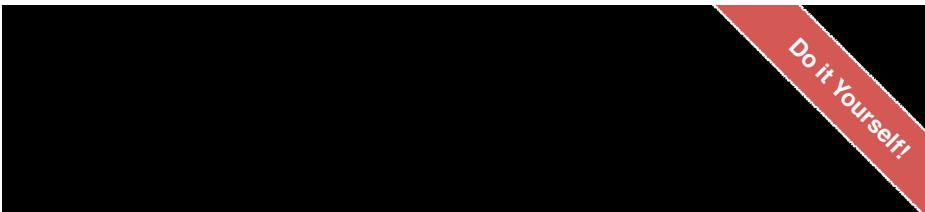


E = Energy
k = constant
D = Dielectric constant (vacuum = 1; H₂O = 80)
q₁ & q₂ = electronic charges (Coulombs)
r = distance (Å)

Hand-on time!

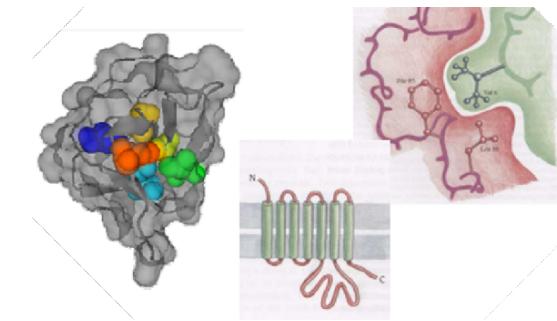
<https://tinyurl.com/bimm143-L11>

Focus on **section 1 to 3** and user your sticky notes for problems and questions please!



Key forces affecting structure:

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



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.

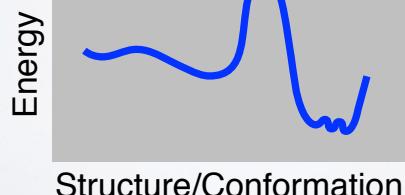
NEXT UP:

- ▶ **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

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

Two main approaches:

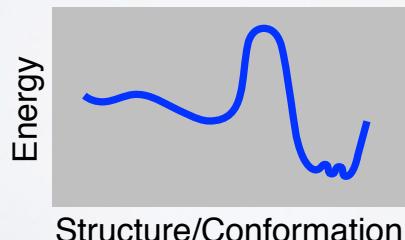
- (1). Physics-Based
- (2). Knowledge-Based



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

Two main approaches:

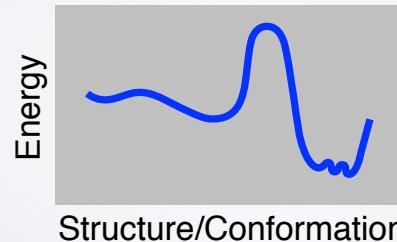
- (1). **Physics-Based**
- (2). Knowledge-Based



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

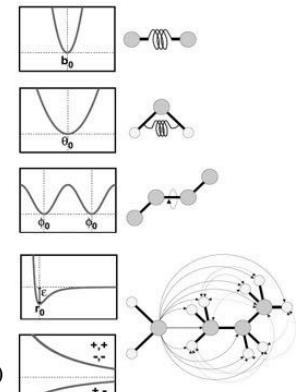
Two main approaches:

- (1). Physics-Based
- (2). Knowledge-Based



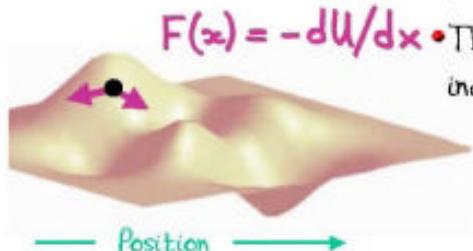
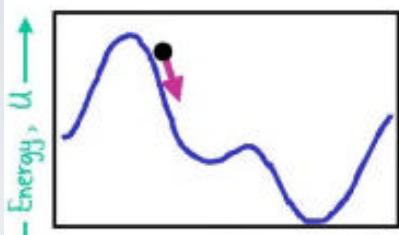
PHYSICS-BASED POTENTIALS
ENERGY TERMS FROM PHYSICAL THEORY

$$U(\vec{R}) = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} + \underbrace{\sum_{dihedrals} k_i^{dih}(1 + \cos(n_i\phi_i + \delta_i))}_{U_{dihedral}} + \underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]}_{U_{nonbond}} + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}$$



CHARMM PE function, see: <http://www.charmm.org/>

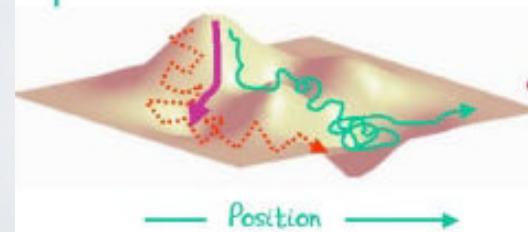
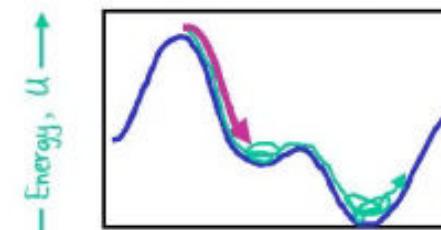
TOTAL POTENTIAL ENERGY



- 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: Bond, Bond angles, Torsion angles and non-bonded atom pairs.

Slide Credit: Michael Levitt

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 $\exp(-\Delta U/kT)$.

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 approxs
Force fields, quantum entropy, water effects
Moore's law: hardware improving

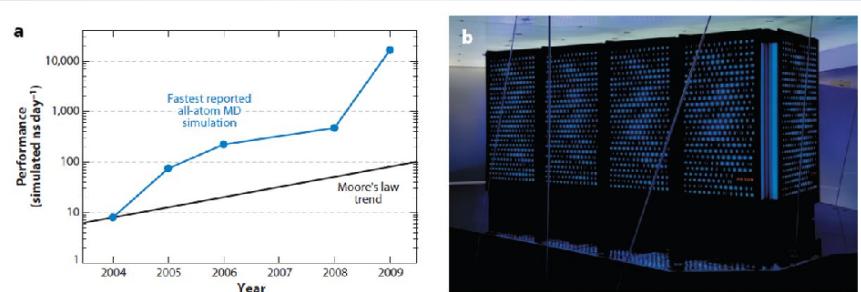
HOW COMPUTERS HAVE CHANGED

DATE	COST	SPEED	MEMORY	SIZE
1967	\$10M	0.1 MHz	1 MB	HALL
2013	\$16,000	1 GHz	10 GB	LAPTOP
CHANGE	10,000	10,000	10,000	10,000

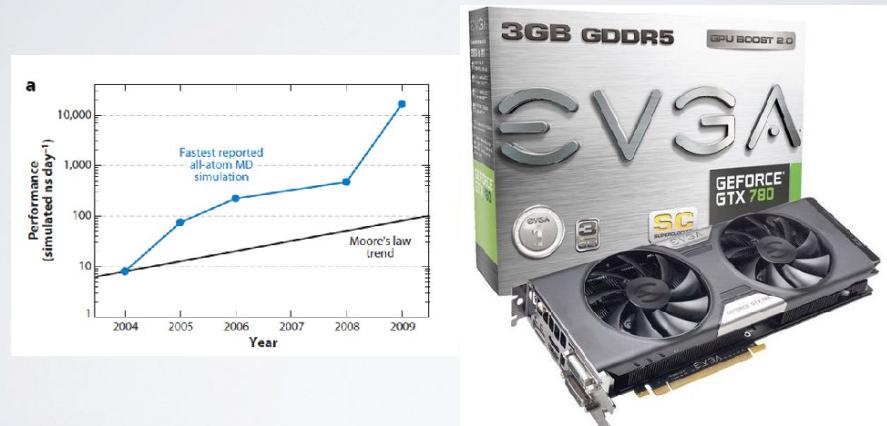
If cars were like computers then a new Vehc
would cost \$3, would have a top speed of
1,000,000 Km/hr, would carry 50,000
adults and would park in a shed.



SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER



SIDE-NOTE: GPUS AND ANTON SUPERCOMPUTER

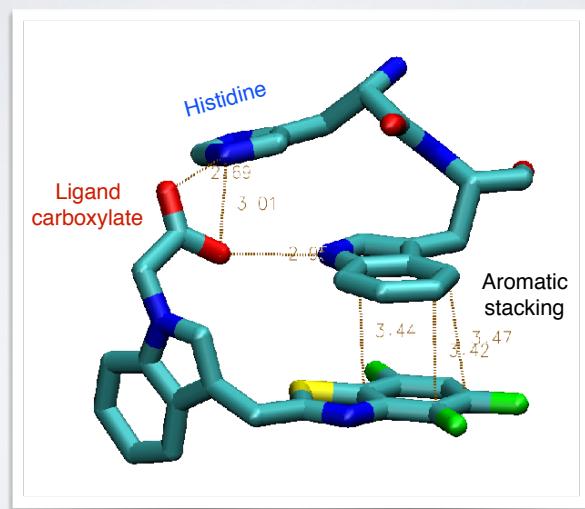


KEY CONCEPT: POTENTIAL FUNCTIONS
DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION
OF ITS **STRUCTURE**

Two main approaches:

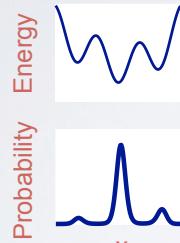
- (1). **Physics-Based**
- (2). **Knowledge-Based**

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

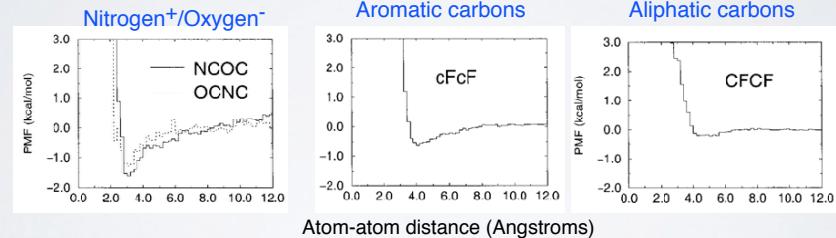
Example: ligand carboxylate O to protein histidine N

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

KNOWLEDGE-BASED DOCKING POTENTIALS

"PMF", Muegge & Martin, J. Med. Chem. (1999) 42:791

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



$$E_{prot-lig} = E_{vdw} + \sum_{pairs(ij)} E_{type(ij)}(r_{ij})$$

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)

Do it Yourself!

Hand-on time!

<https://tinyurl.com/bimm143-L11>

Focus on **section 4 & 5**

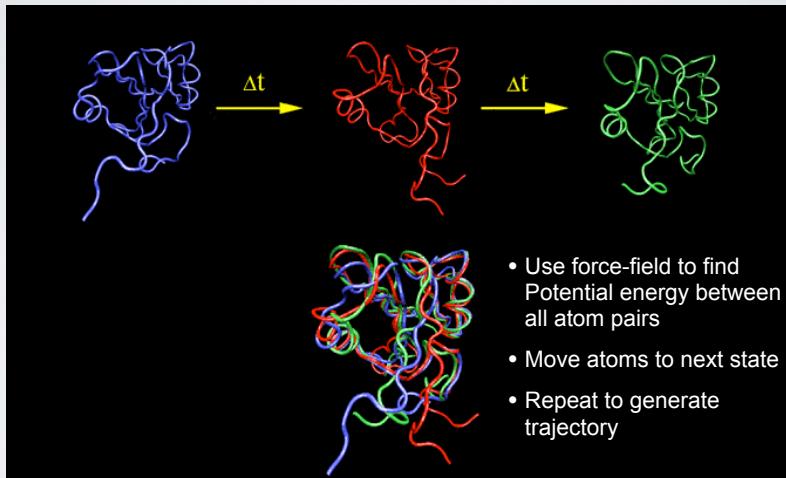
NEXT UP:

- ▶ 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



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

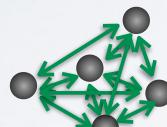
- ▶ Divide **time** into discrete ($\sim 1\text{fs}$) **time steps** (Δt)
(for integrating equations of motion, see below)



- 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

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

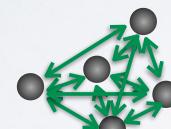
Empirical force field

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

- 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)



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

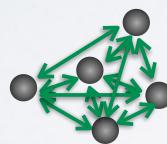
$$r(t + \Delta t) = r(t) + v(t + \frac{\Delta t}{2}) \Delta t$$

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 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)

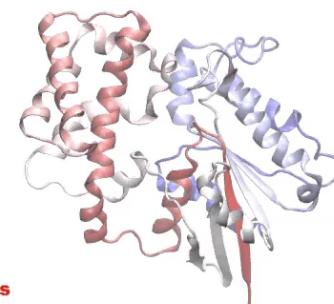


$$v(t + \frac{\Delta t}{2}) = v(t - \frac{\Delta t}{2}) + \frac{F(t)}{m} \Delta t$$

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

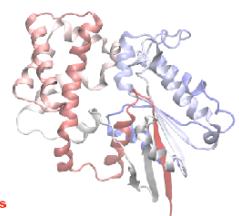
MD Prediction of Functional Motions

Accelerated MD simulation of nucleotide-free transducin alpha subunit

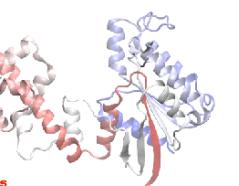


0.00 ns

"close"

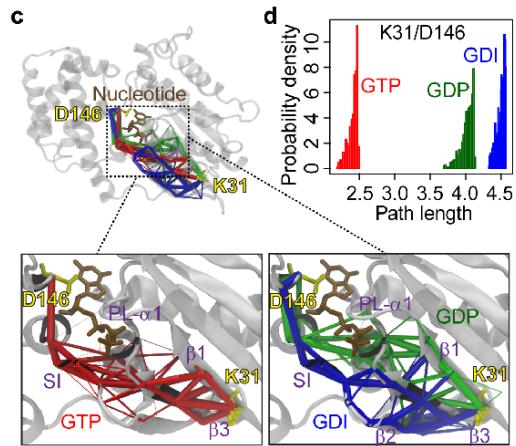


"open"



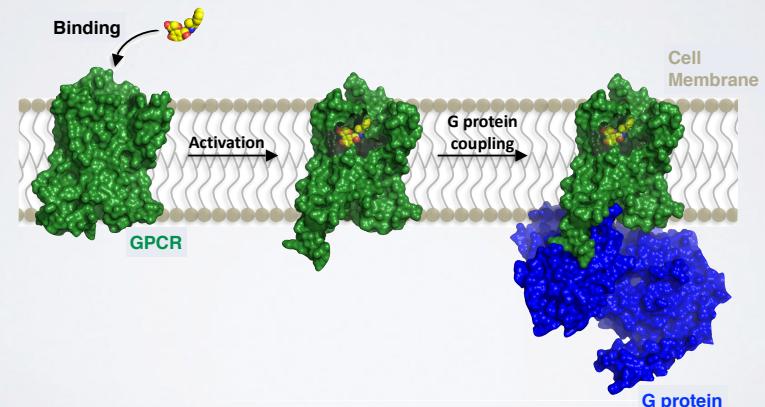
Yao and Grant, Biophys J. (2013)

Simulations Identify Key Residues Mediating Dynamic Activation

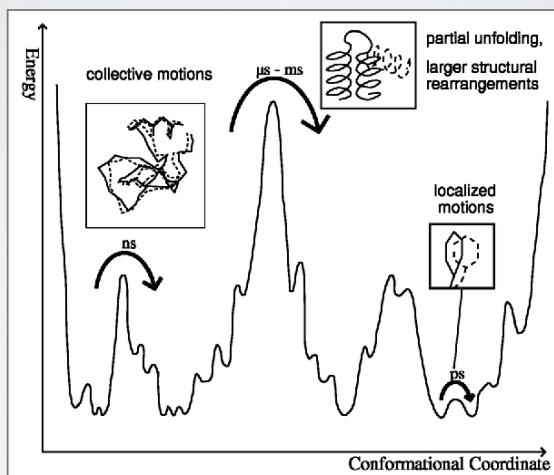


Yao ... Grant, *Journal of Biological Chemistry* (2016)

EXAMPLE APPLICATION OF MOLECULAR SIMULATIONS TO GPCRS



PROTEINS JUMP BETWEEN MANY, HIERARCHICALLY ORDERED “CONFORMATIONAL SUBSTATES”



H. Frauenfelder et al., *Science* **229** (1985) 337

MOLECULAR DYNAMICS IS VERY

Improve this slide

Example: F₁-ATPase in water (183,674 atoms) for 1 nanosecond:

=> 10⁶ integration steps
=> 8.4 * 10¹¹ floating point operations/step
[n(n-1)/2 interactions]

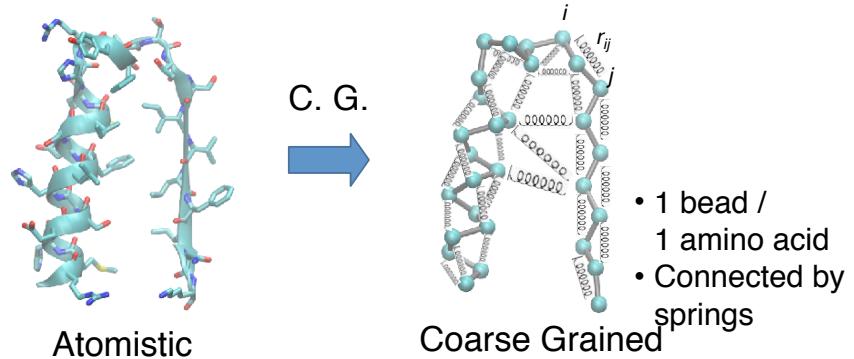
Total: 8.4 * 10¹⁷ flop
(on a 100 Gflop/s cpu: **ca 25 years!**)

... but performance has been improved by use of:

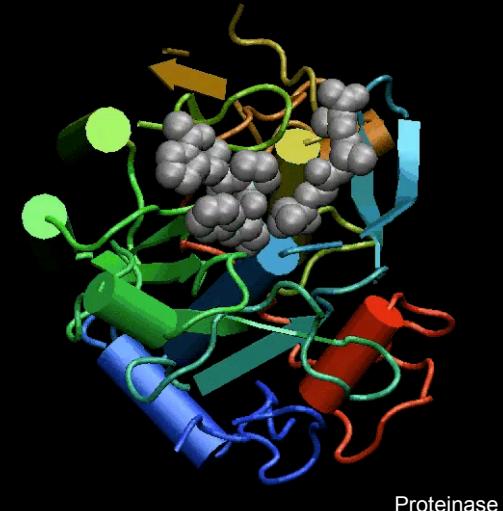
multiple time stepping	ca. 2.5 years
fast multipole methods	ca. 1 year
parallel computers	ca. 5 days
modern GPUs	ca. 1 day
(Anton supercomputer	ca. minutes)

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.



NMA models the protein as a network of elastic strings



Proteinase K

Hand-on time!

<https://tinyurl.com/bimm143-L11>

Focus on **section 6** to **7**

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