

“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

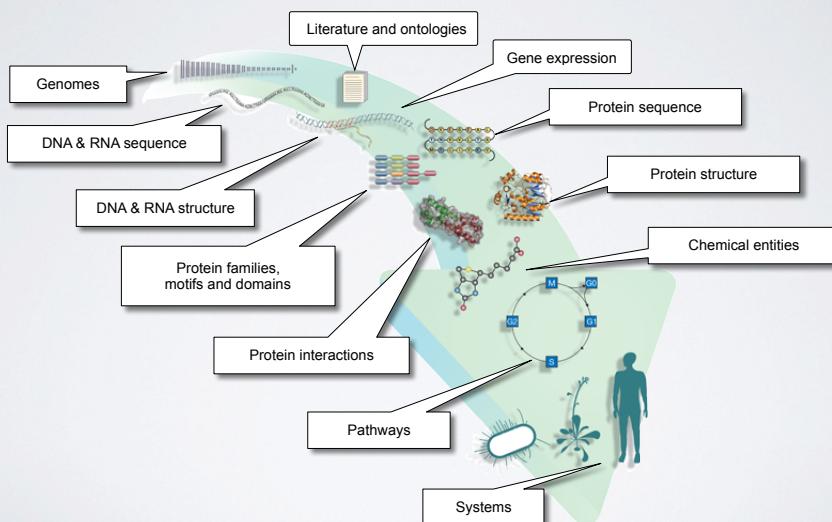
Why should we care?

Why should we care?

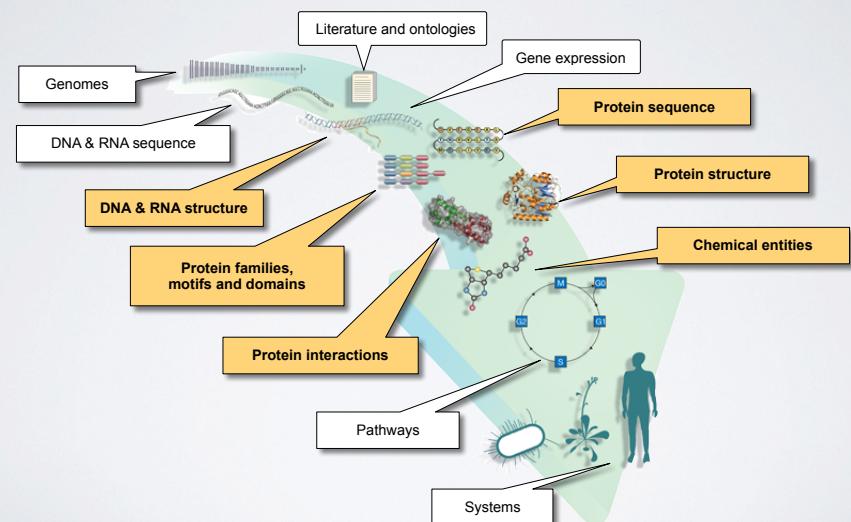
Because biomolecules are “nature’s robots”

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

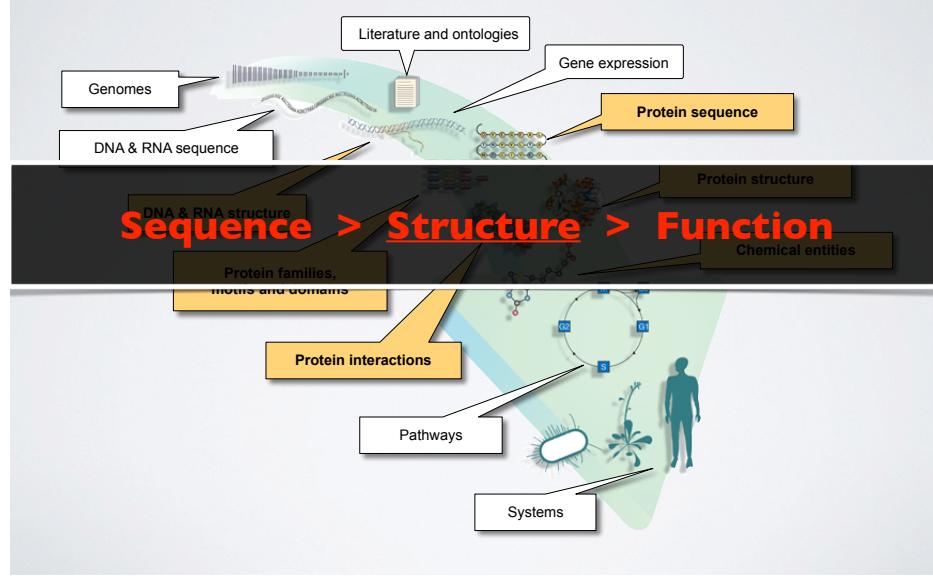
BIOINFORMATICS DATA



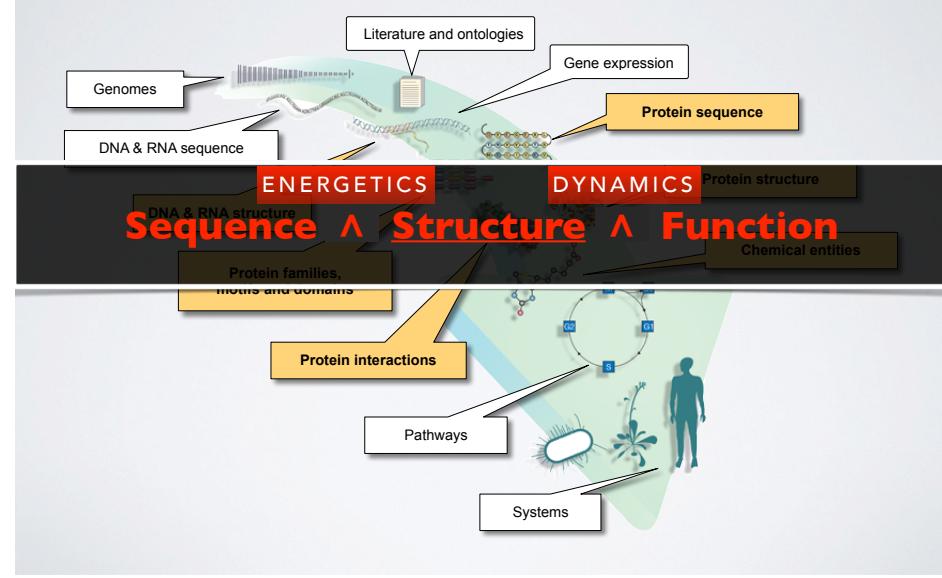
STRUCTURAL DATA IS CENTRAL

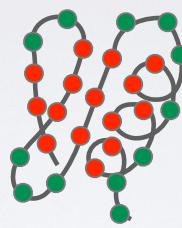
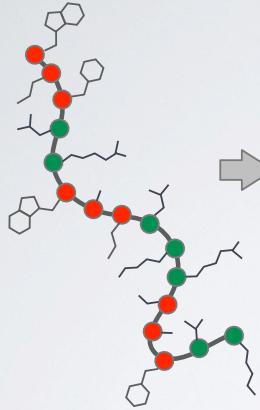


STRUCTURAL DATA IS CENTRAL



STRUCTURAL DATA IS CENTRAL





Sequence

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

Structure

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

Function

- Active in specific "conformations"
- Specific associations & precise reactions

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



Genomics is a great start

Track Bike – DL 175

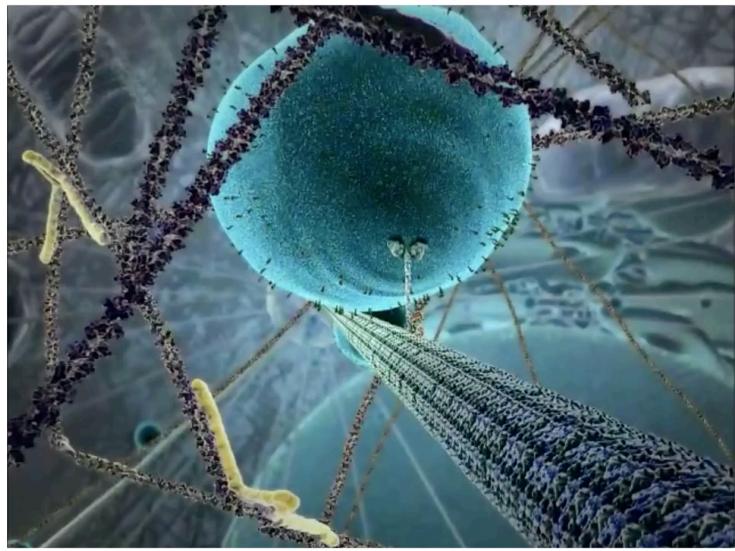
REF. NO.	IBM NO.	DESCRIPTION
1	156011	Track Frame 21", 22", 23", 24", Team Red
2	157040	Fork for 21" Frame
2	157039	Fork for 22" Frame
2	157038	Fork for 23" Frame
2	157037	Fork for 24" Frame
3	191202	Handlebar TTT Competition Track Alloy 15/16" Headset Stem, TTT, Specify extension
4	145937	Examiner Bolt
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	145834	Fixed Belt
20	145835	Fixing Washer
21	145822	Dustcap
22	145823	R.H. and L.H. Crankset with Chainwheel
23	146472	Fixed Cup
24	146472	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

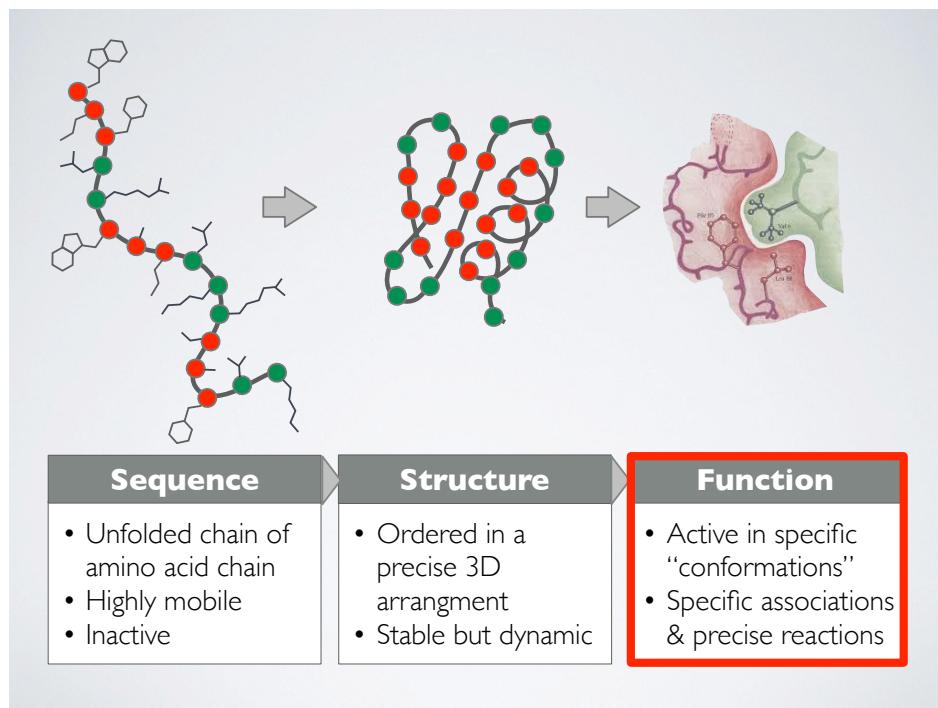
... but not the end



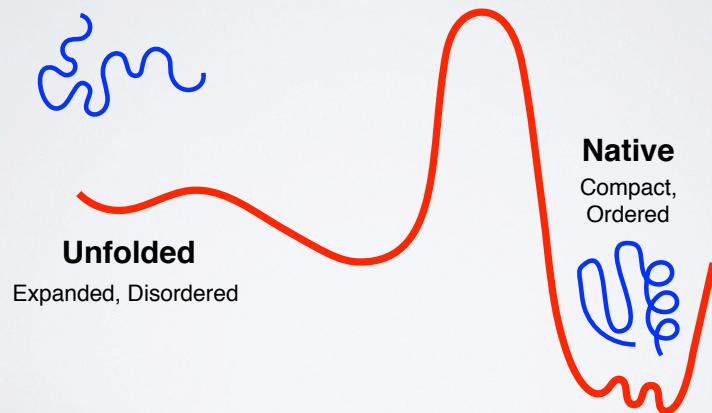
- We want the full spatiotemporal picture, and an ability to control it
- Broad applications, including drug design, medical diagnostics, chemical manufacturing, and energy



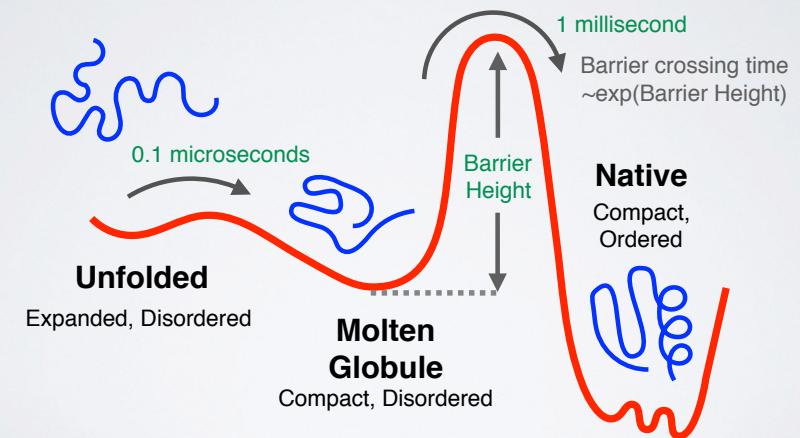
Extracted from The Inner Life of a Cell by Cellular Visions and Harvard
[YouTube link: <https://www.youtube.com/watch?v=y-uuk4Pr2i8>]



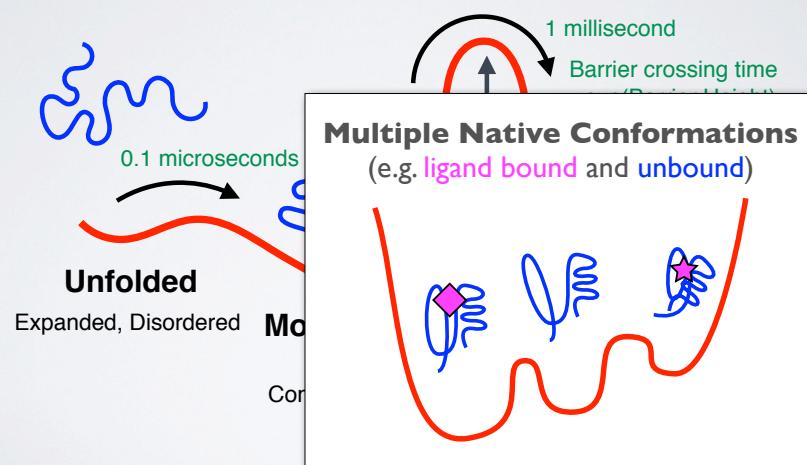
KEY CONCEPT: ENERGY LANDSCAPE



KEY CONCEPT: ENERGY LANDSCAPE



KEY CONCEPT: ENERGY LANDSCAPE



Today's Menu

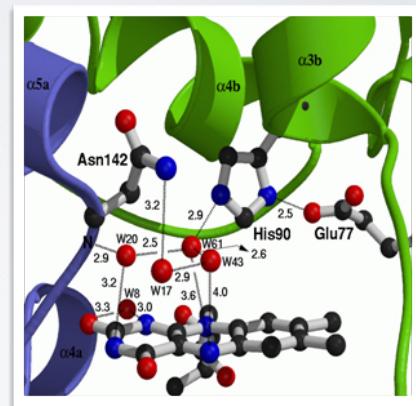
- Overview of structural bioinformatics
 - Motivations, goals and challenges
- Fundamentals of protein structure
 - Structure composition, form and forces
- Representing, interpreting & modeling protein structure
 - Visualizing & interpreting protein structures
 - Analyzing protein structures
 - Modeling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

Today's Menu

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

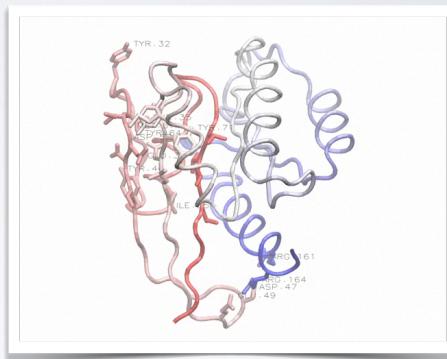
Motivation 1: Detailed understanding of molecular interactions

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



Motivation 1: Detailed understanding of molecular interactions

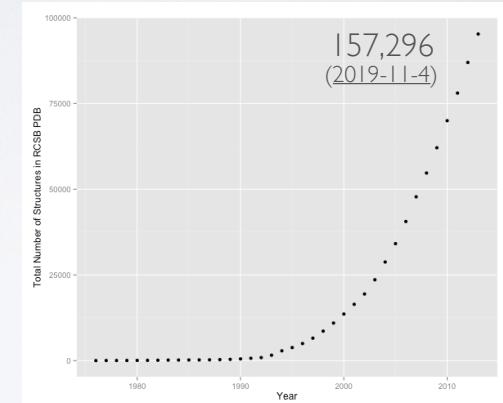
Computational modeling can provide detailed insight into functional interactions, their regulation and potential consequences of perturbation.



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: <https://www.rcsb.org/stats/>

Motivation 2: Lots of structural data is becoming available

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

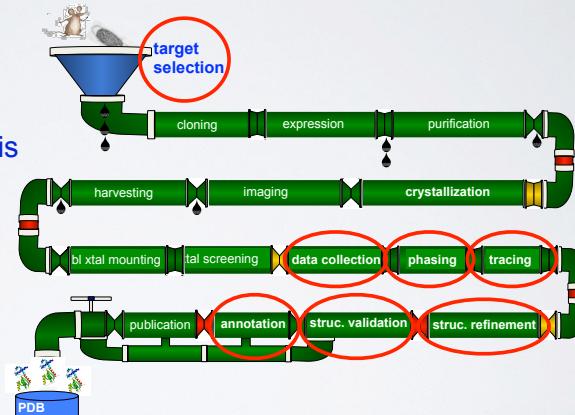
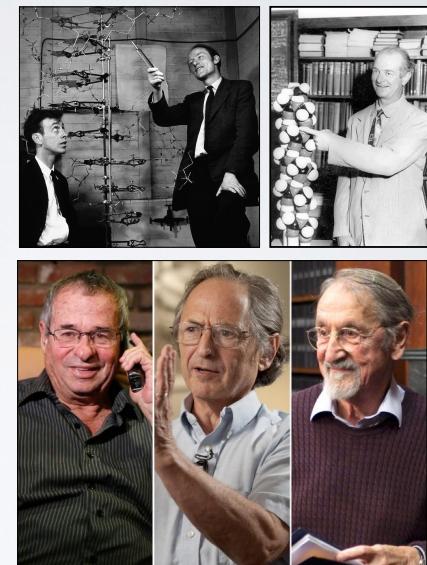


Image Credit: "Structure determination assembly line" Adam Godzik

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



SUMMARY OF KEY **MOTIVATIONS**

Sequence > Structure > Function

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

Structure is more conserved than sequence

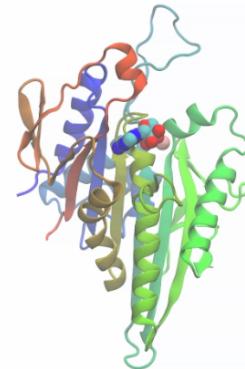
- Structure allows identification of more distant evolutionary relationships

Structure is encoded in sequence

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

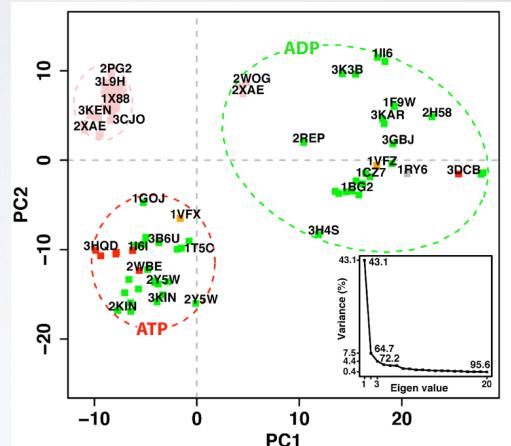
Goals:

- Visualization
- Analysis
- Comparison
- Prediction
- Design



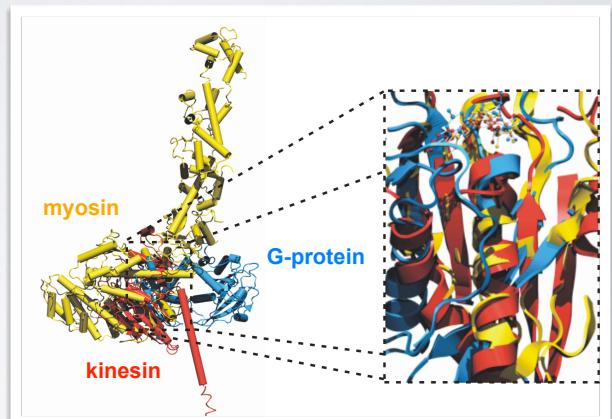
Scarabelli and Grant. PLoS. Comp. Biol. (2013)

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



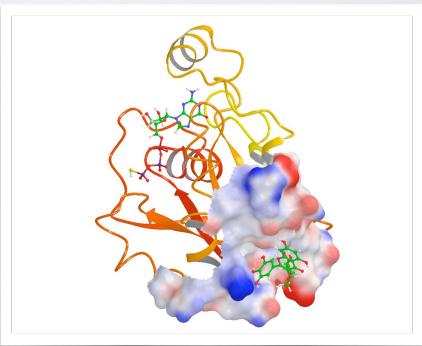
Scarabelli and Grant. PLoS. Comp. Biol. (2013)

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



Grant et al. unpublished

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



Grant et al. PLoS One (2011, 2012)

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



Grant et al. PLoS Biology (2011)

MAJOR RESEARCH AREAS AND CHALLENGES

Include but are not limited to:

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

With applications to Biology, Medicine, Agriculture and Industry

Today's Menu

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

HIERARCHICAL STRUCTURE OF PROTEINS

Primary > Secondary > Tertiary > Quaternary

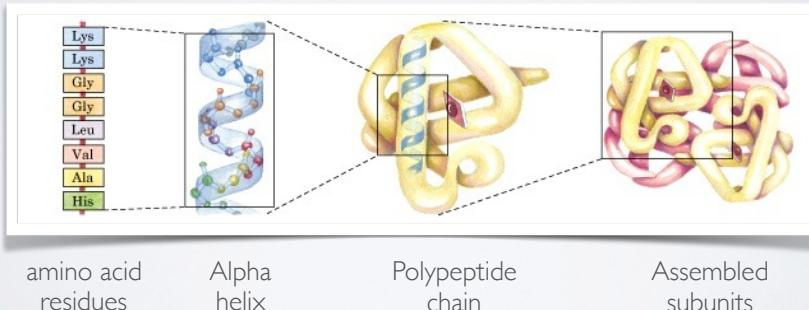


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

RECAP: AMINO ACID NOMENCLATURE

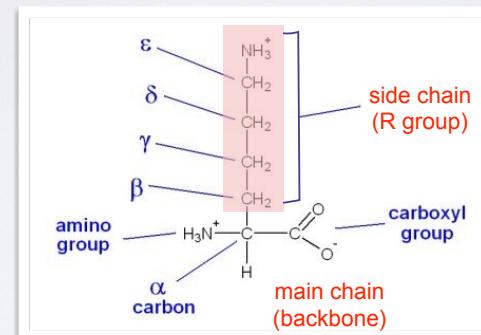


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

AMINO ACIDS CAN BE GROUPED BY THE PHYSIOCHEMICAL PROPERTIES

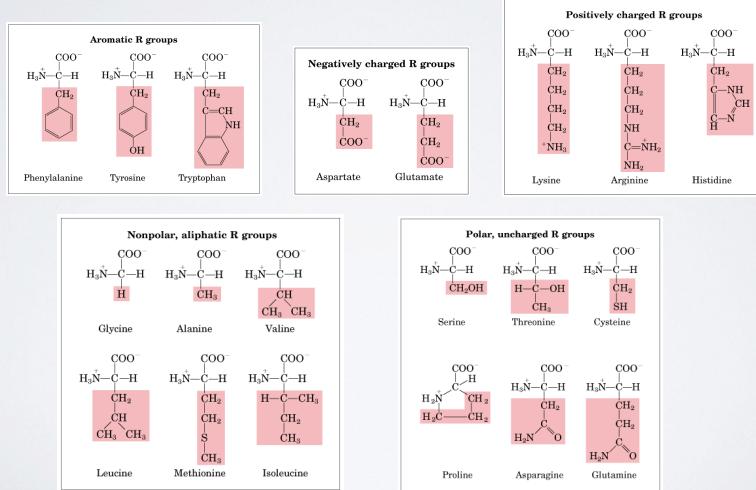


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

AMINO ACIDS POLYMERIZE THROUGH PEPTIDE BOND FORMATION

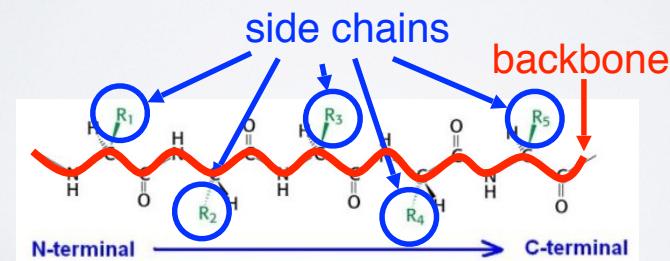
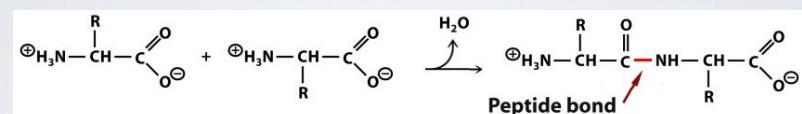
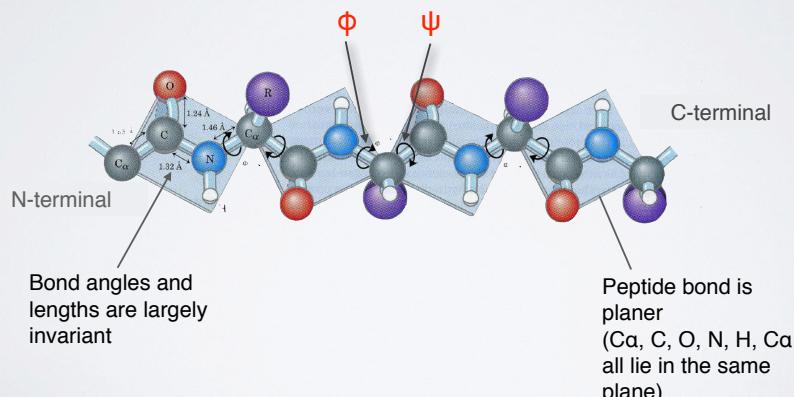
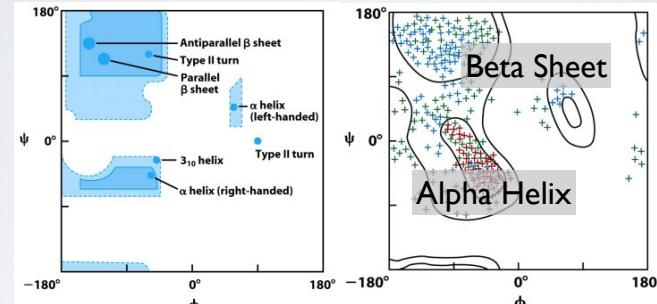


Image from: <http://www.ncbi.nlm.nih.gov/books/NBK21581/>

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

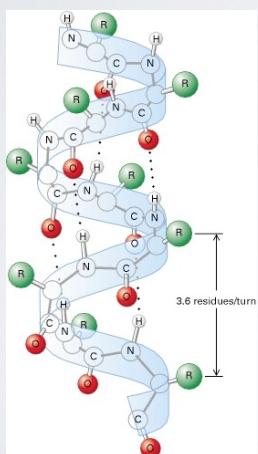


PHI VS PSI PLOTS ARE KNOWN AS RAMACHANDRAN DIAGRAMS



- 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

MAJOR SECONDARY STRUCTURE TYPES ALPHA HELIX & BETA SHEET

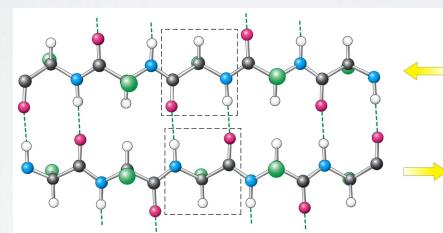


α -helix

- Most common form 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
- 3₁₀-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

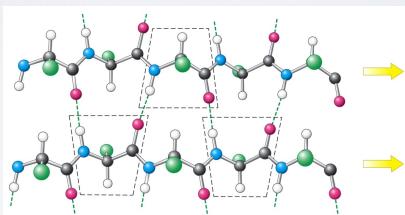


In antiparallel β -sheets

- 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

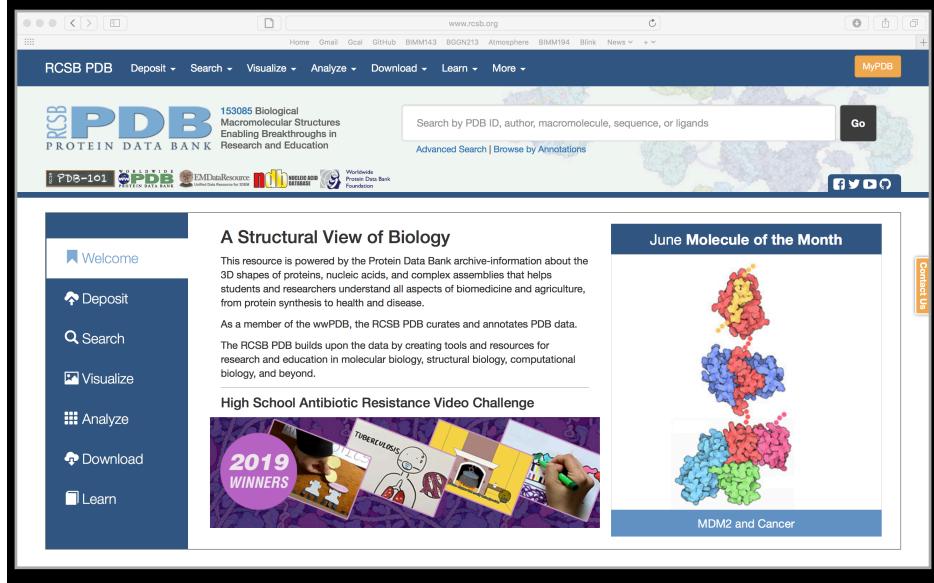


In parallel β -sheets

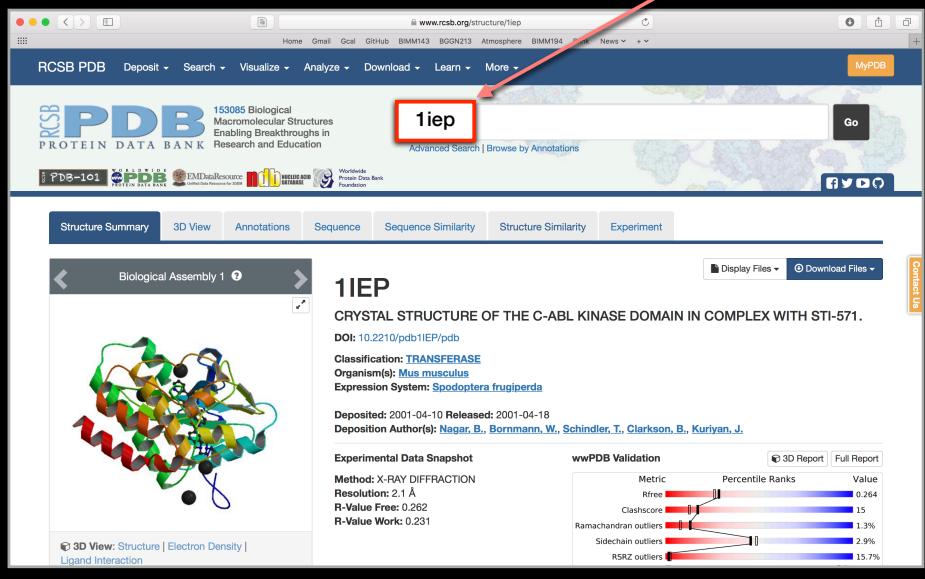
- Adjacent β -strands run in same direction
- Hydrogen bonds (dashed lines) between NH and CO stabilize the structure
- The side chains (in green) are above and below the sheet

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

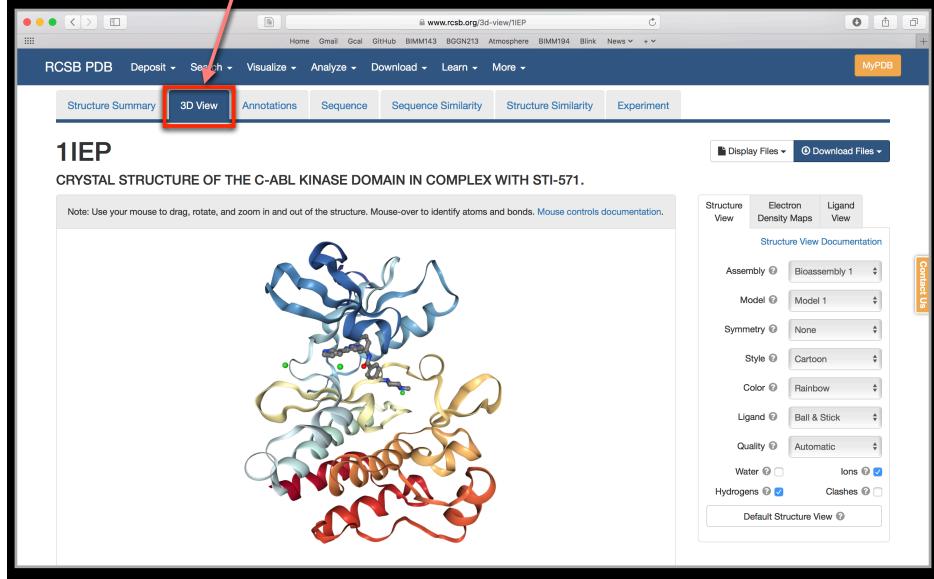
Protein Data Bank (PDB) is the main repository for Biomolecular structure data
<http://www.rcsb.org>

The screenshot shows the RCSB PDB homepage. The top navigation bar includes links for Home, Gmail, Gcal, GitHub, BIMM143, BGGN213, Atmosphere, BIMM194, Blink, News, and More. The main content area features the RCSB PDB logo and a search bar. A sidebar on the left lists Welcome, Deposit, Search, Visualize, Analyze, Download, and Learn. A "June Molecule of the Month" section highlights the MDM2 and Cancer complex. A "High School Antibiotic Resistance Video Challenge" banner for 2019 winners is also visible.

You can search by text (e.g. "ABL kinase"), PDB code (e.g. "[1iep](http://www.rcsb.org/structure/1iep)") or sequence
<http://www.rcsb.org>

The screenshot shows the search results for PDB code 1iep. The URL in the address bar is http://www.rcsb.org/structure/1iep. The page title is "1IEP". Below it, the text reads "CRYSTAL STRUCTURE OF THE C-ABL KINASE DOMAIN IN COMPLEX WITH STI-571.". The classification is listed as TRANSFERASE, and the organism is Mus musculus. The experimental data snapshot table includes columns for Method (X-RAY DIFFRACTION), Resolution (2.1 Å), R-Value Free (0.262), and R-Value Work (0.231). A 3D ribbon model of the protein structure is displayed at the bottom.

You can get a **3D View** of and read details about the experiment and molecule
<http://www.rcsb.org>

The screenshot shows the 3D view of structure 1IEP. The URL in the address bar is http://www.rcsb.org/3d-view/1IEP. The page title is "1IEP". The main content area displays a 3D ribbon model of the protein domain. On the right, there is a panel for "Structure View Documentation" with various settings for assembly, model, symmetry, style, color, ligand, quality, water, ions, hydrogens, and clashes. A note at the top says, "Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Mouse-over to identify atoms and bonds. [Mouse controls documentation](#)."

You can display or download PDB format files for a particular entry

<http://www.rcsb.org>

1IEP
CRYSTAL STRUCTURE OF THE C-ABL KINASE DOMAIN IN COMPLEX WITH STI-571.

Note: Use your mouse to drag, rotate, and zoom in and out of the structure. Mouse-over to identify atoms and bonds. [Mouse controls documentation](#).

Structure View Electron Density Maps Ligand View

Assembly Biossembly 1 Model Model 1 Symmetry None Style Cartoon Color Rainbow Ligand Ball & Stick Quality Automatic Water Ions Hydogen Clashes Default Structure View

Side-Note: PDB File Format

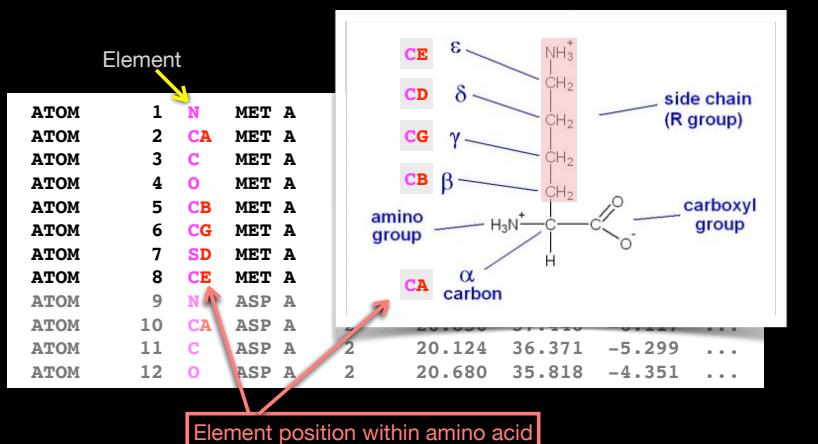
- PDB files contains atomic **coordinates** and associated information.

Element	Amino Acid	Sequence/Residue Number	Coordinates			(etc.)
			X	Y	Z	
ATOM	1 N	MET A	1	19.353	41.547	-3.887
ATOM	2 CA	MET A	1	20.513	40.939	-4.592
ATOM	3 C	MET A	1	20.150	39.658	-5.355
ATOM	4 O	MET A	1	19.053	39.551	-5.903
ATOM	5 CB	MET A	1	21.642	40.678	-3.592
ATOM	6 CG	MET A	1	21.233	39.903	-2.360
ATOM	7 SD	MET A	1	22.533	39.928	-1.113
ATOM	8 CE	MET A	1	23.771	38.881	-1.885
ATOM	9 N	ASP A	2	21.068	38.694	-5.390
ATOM	10 CA	ASP A	2	20.856	37.440	-6.117
ATOM	11 C	ASP A	2	20.124	36.371	-5.299
ATOM	12 O	ASP A	2	20.680	35.818	-4.351

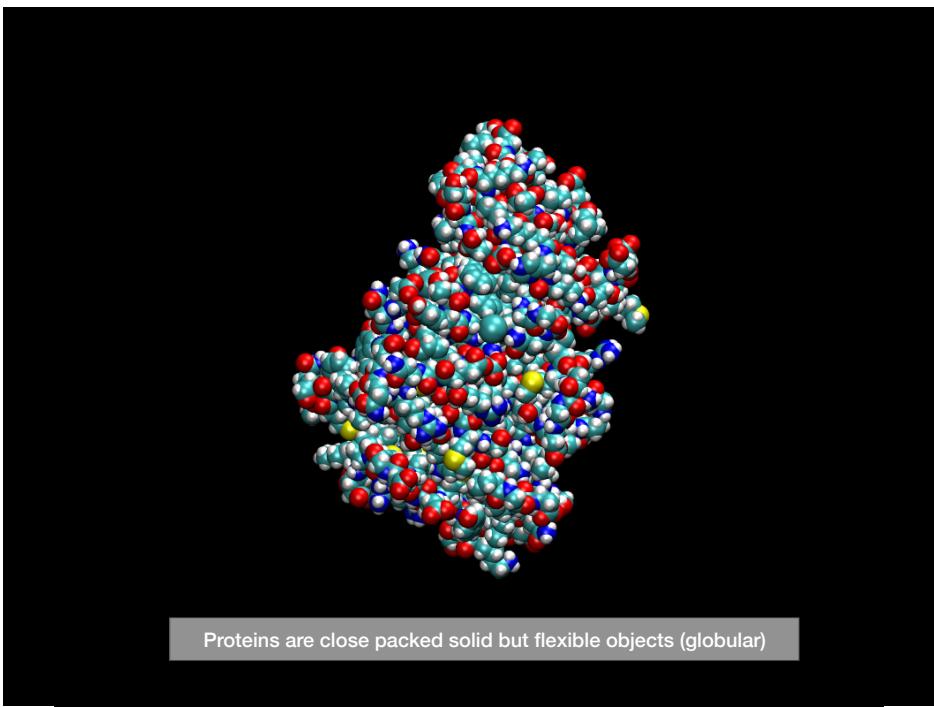
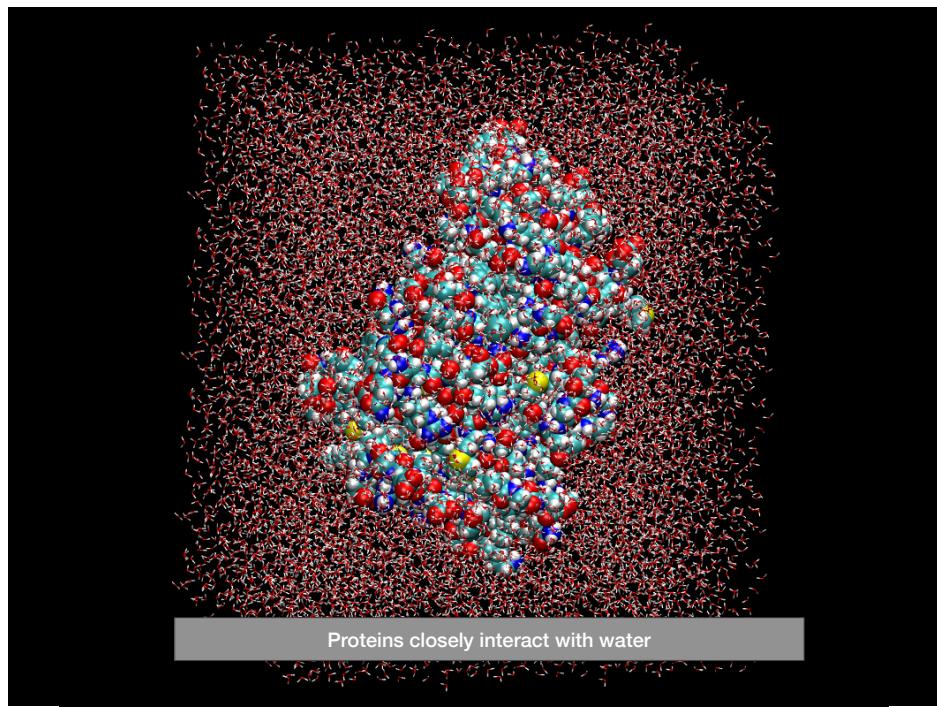
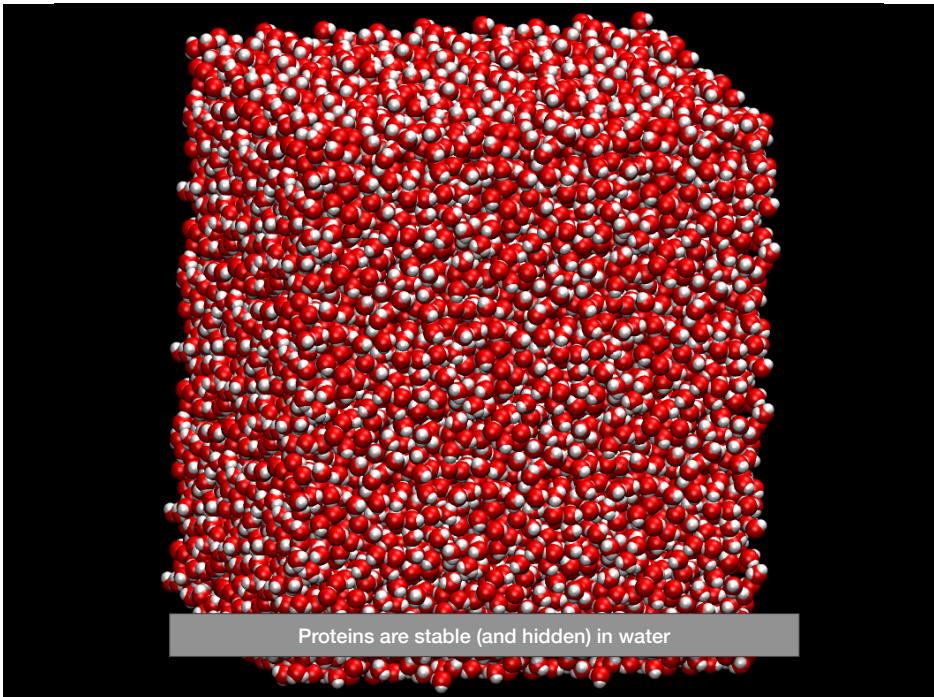
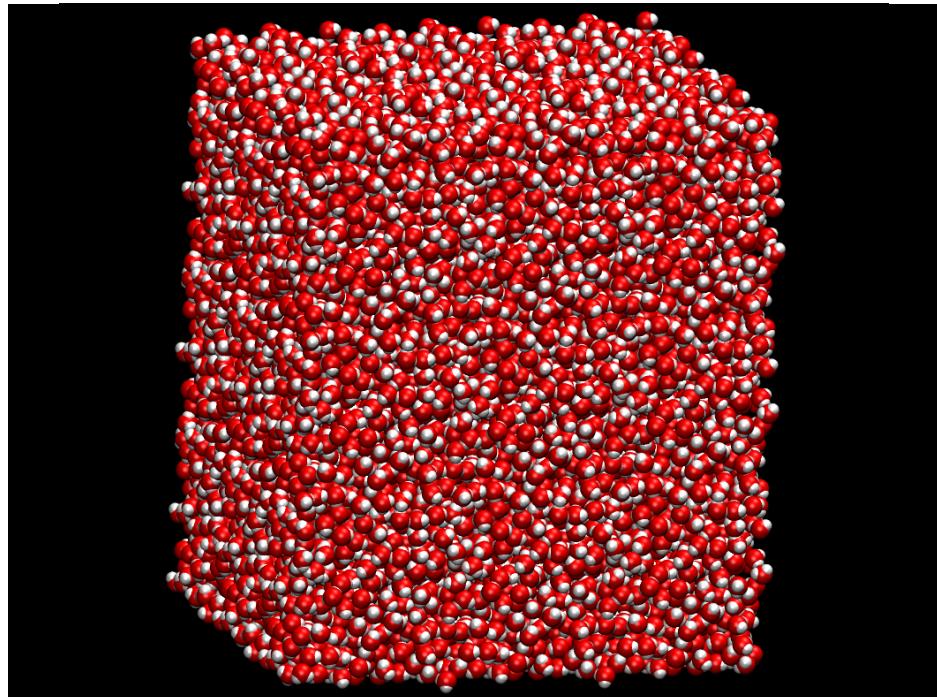
Element position within amino acid

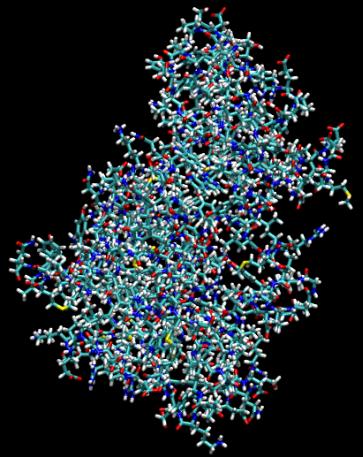
Side-Note: PDB File Format

- PDB files contains atomic **coordinates** and associated information.

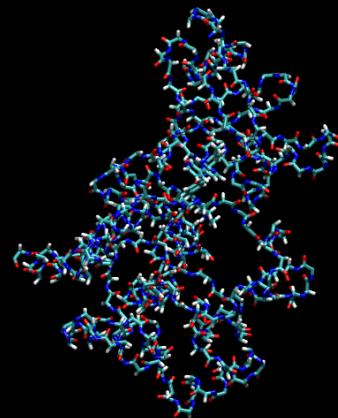


What Does a Protein Look like?

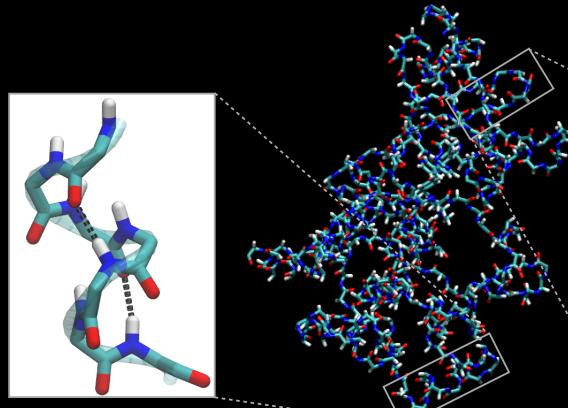




Due to their large size and complexity it is often hard to see what's important in the structure



Backbone or main-chain representation can help trace chain topology



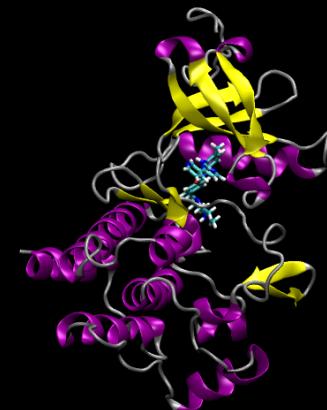
Backbone or main-chain representation can help trace chain topology & reveal secondary structure



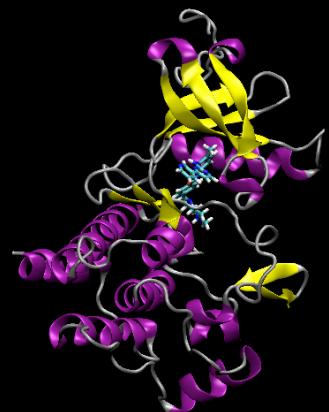
Tube or trace representation is one of the simplest views



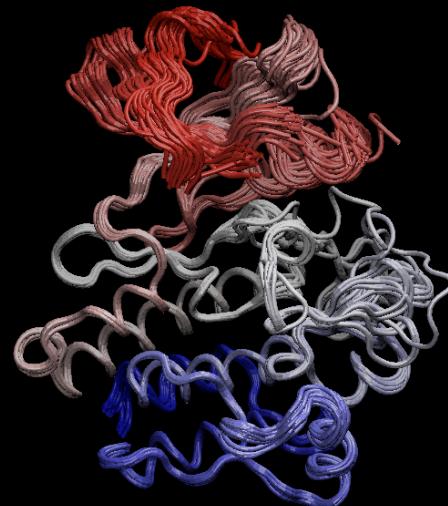
Tube with added colors to highlight **secondary structure**



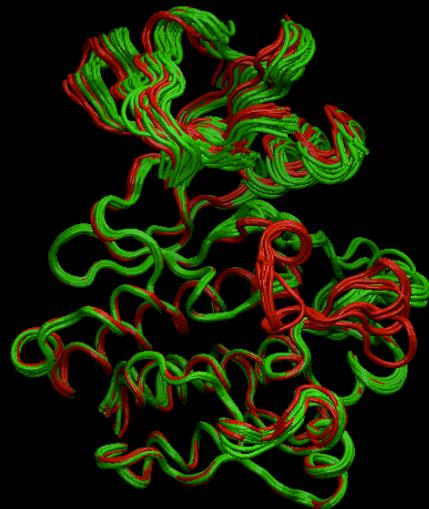
Simplified "cartoon" secondary structure representations are commonly used to communicate structural details



Viewing in 3D is often essential for interpretation.
Now we can clearly see 2^o and 3^o structure - the coiled chain of connected secondary structures



Viewing multiple superposed structures solved under different conditions can highlight **flexible regions**

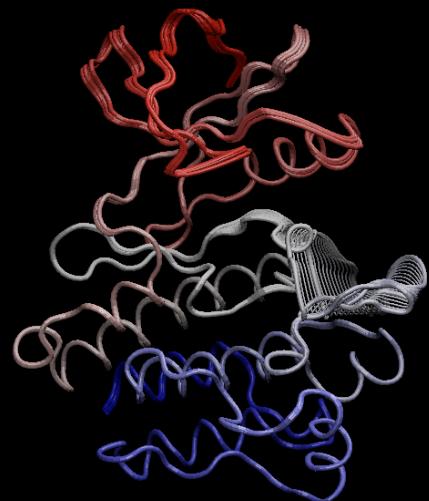


Active
Inactive

Viewing multiple superposed structures solved under different conditions can highlight **distinct conformations**



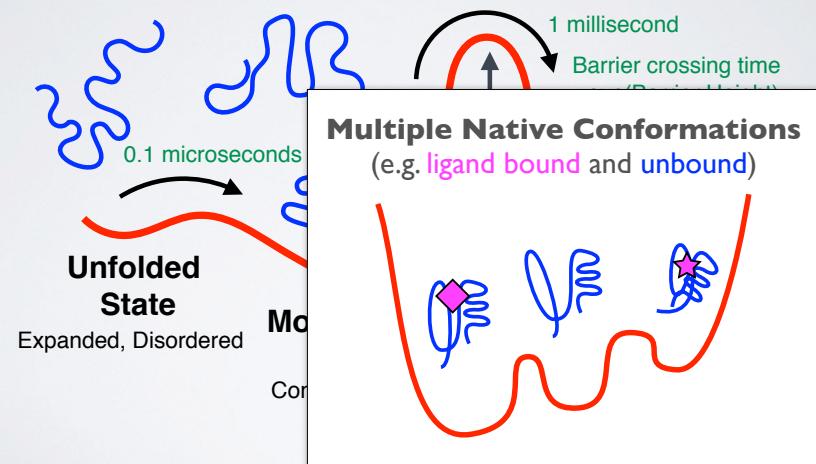
Analyzing these multiple structures can reveal **functional motions**
- i.e. displacements that are essential for regulating function



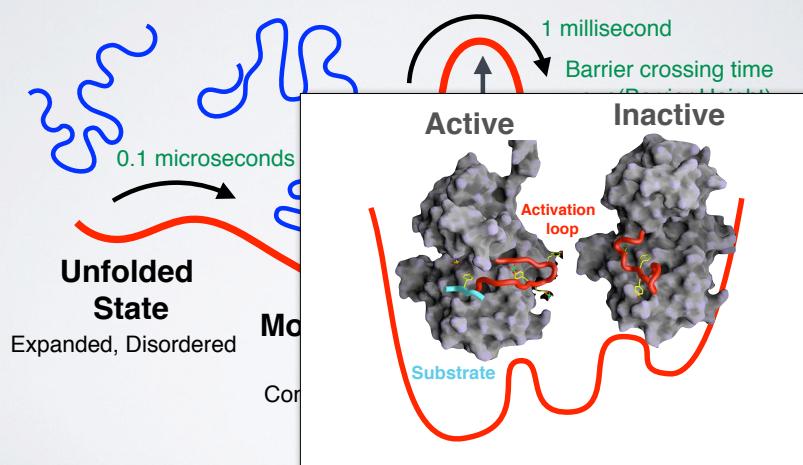
"Activation loop"

Analyzing these multiple structures can reveal **functional motions**
- i.e. displacements that are essential for regulating function

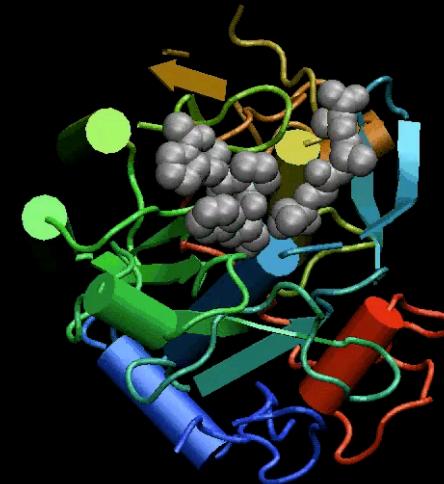
KEY CONCEPT: ENERGY LANDSCAPE



KEY CONCEPT: ENERGY LANDSCAPE



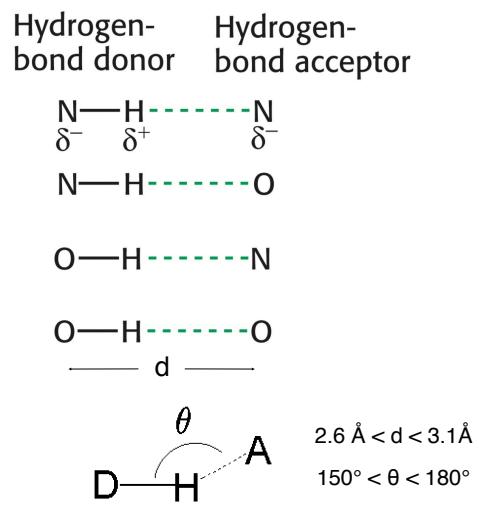
Normal Mode Analysis (NMA) models the protein as a network of elastic strings



NMA is a bioinformatics method to predict the intrinsic dynamics of biomolecules

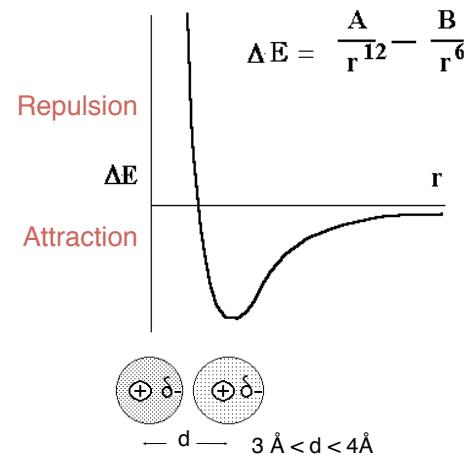
Key forces affecting structure:

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



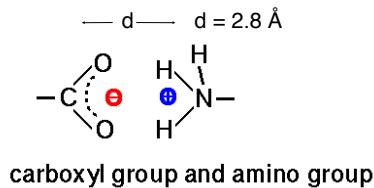
Key forces affecting structure:

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

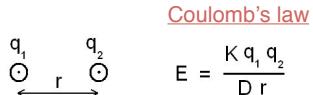


Key forces affecting structure:

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



(some time called IONIC BONDS or SALT BRIDGEs)



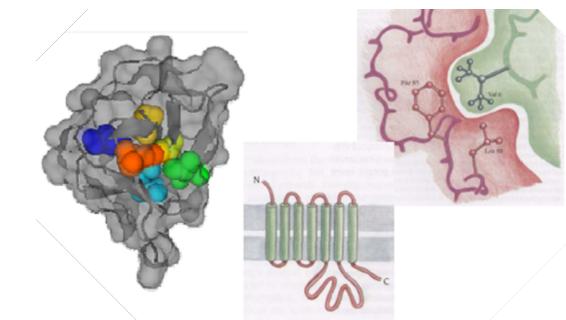
Coulomb's law

$$E = \frac{K q_1 q_2}{D r}$$

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

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.

Today's Menu

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

Today's Menu

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

Do it Yourself!

Hand-on time!

Focus on **section 1** only please!

[~20 mins]

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

Side-Note: PDB File Format

- PDB files contains atomic **coordinates** and associated information.

ATOM	Element	Amino Acid	Chain	Sequence/Residue Number	Coordinates			(etc.)
					X	Y	Z	
ATOM	1	N	MET	A	1	19.353	41.547	-3.887
ATOM	2	CA	MET	A	1	20.513	40.939	-4.592
ATOM	3	C	MET	A	1	20.150	39.658	-5.355
ATOM	4	O	MET	A	1	19.053	39.551	-5.903
ATOM	5	CB	MET	A	1	21.642	40.678	-3.592
ATOM	6	CG	MET	A	1	21.233	39.903	-2.360
ATOM	7	SD	MET	A	1	22.533	39.928	-1.113
ATOM	8	CE	MET	A	1	23.771	38.881	-1.885
ATOM	9	N	ASP	A	2	21.068	38.694	-5.390
ATOM	10	CA	ASP	A	2	20.856	37.440	-6.117
ATOM	11	C	ASP	A	2	20.124	36.371	-5.299
ATOM	12	O	ASP	A	2	20.680	35.818	-4.351

Element position within amino acid

Side-Note: PDB File Format

- PDB files contains atomic **coordinates** and associated information.

ATOM	Element							
ATOM	1	N	MET	A				
ATOM	2	CA	MET	A				
ATOM	3	C	MET	A				
ATOM	4	O	MET	A				
ATOM	5	CB	MET	A				
ATOM	6	CG	MET	A				
ATOM	7	SD	MET	A				
ATOM	8	CE	MET	A				
ATOM	9	N	ASP	A				
ATOM	10	CA	ASP	A				
ATOM	11	C	ASP	A				
ATOM	12	O	ASP	A				

Element position within amino acid

Download VMD

Hands-on Time!

Focus on **section 2** of "Lab Sheet" (using VMD)
[~30 mins]

Today's Menu

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

Do it Yourself!

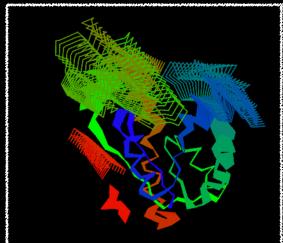
Hand-on time!

Focus on **section 3 to 5**

[~60 mins]

Bio3D view()

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



- In your R console:

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

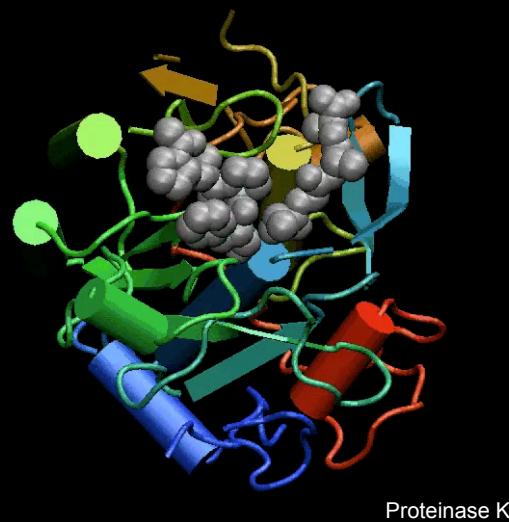
- To use in your R session:

```
> library("bio3d.view")
> pdb <- read.pdb("5p21")
> view(pdb)
> view(pdb, "overview", col="sse")
```

Today's Menu

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

NMA models the protein as a network of elastic strings



Proteinase K

NMA in Bio3D

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

```
```{r}
library(bio3d)
library(bio3d.view)
...````
```

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

Bio3D view()

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

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

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

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

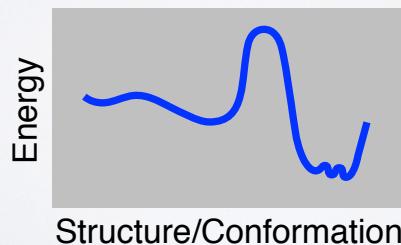
Two main approaches:

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

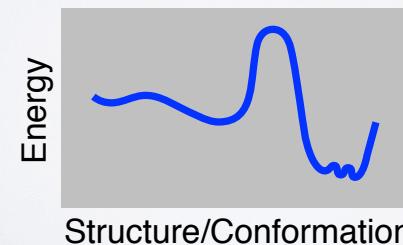
KEY CONCEPT: POTENTIAL FUNCTIONS
DESCRIBE A SYSTEMS **ENERGY** AS A FUNCTION
OF ITS **STRUCTURE**

Two main approaches:

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



This will be the focus of the next class!



SUMMARY

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

[[Muddy Point Assessment](#)]