Exercises

1) Give three advantages of using Biopython? parse common file formats, interact with biological databases, interface with other software
2) What are three ways of accessing Ensembl, which would you use for single locus query, groups of genes, whole genomes? website (single locus), REST/Biomart (groups), ftp/perl API (genomes)
4) Why use pipes in the command line interface?
eliminate intermediate files, save space
5) What is the advantage of using bash or Python to create a bioinformatics pipeline (combination of commands to process data)? parallel processing, reproducibility, handle multiple files or multiple parameters combinations, optimize performance
6) Why is API a better way to get information from a database than (i) web-browser, (ii) download database? web-browser only gives one gene at a time, database download can be large and is not specialized information
7) What is the advantage of shotgun metagenomic compared to 16S PCR microbiome analysis? examines bacteria, fungal, viral at the same time, no PCR bias, strain level resolution, functional analysis by gene content
8) How would you compare the function of two microbiomes? GO/KEGG analysis

Today's objectives

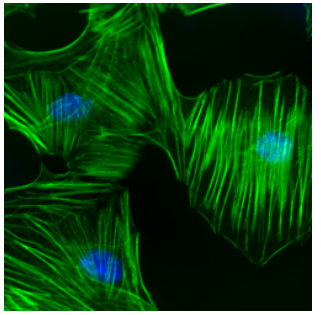
- Protein structure and function
- Protein structure prediction
- BioPhysics, homology modeling, evolution
- Predicting changes in structure

Areas within computational biology

Genomics

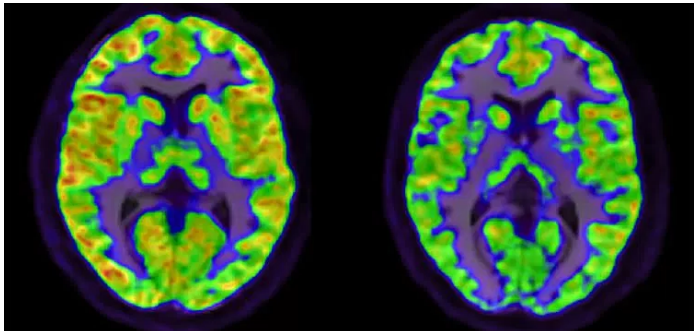
```
GCACGATCGATCA
GCACGAACCATCA
GCA-GATCCATCA
GCA-GATCGATCA
```

Cell biology & Development

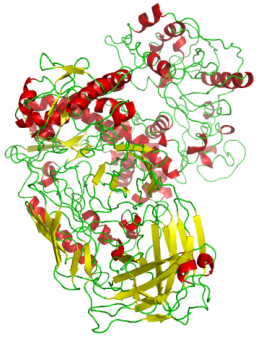


Protein structures
Biophysics

Neuroscience

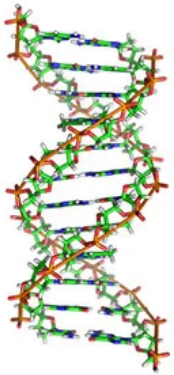


Biochemistry

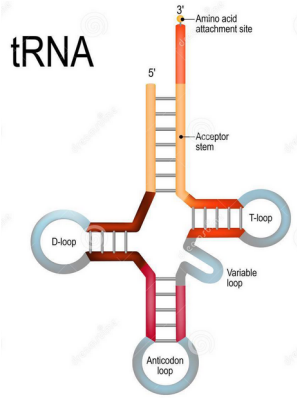


-new concepts
-shared methods

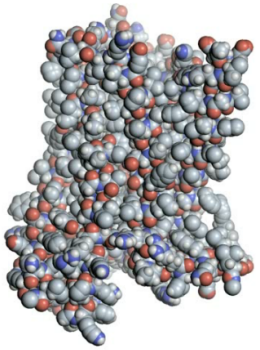
Structures



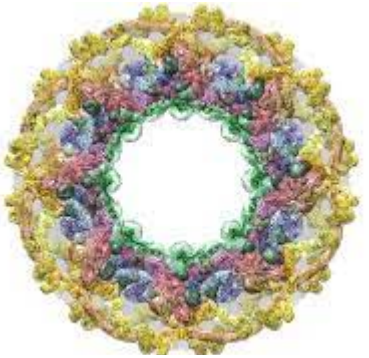
DNA



RNA



Protein



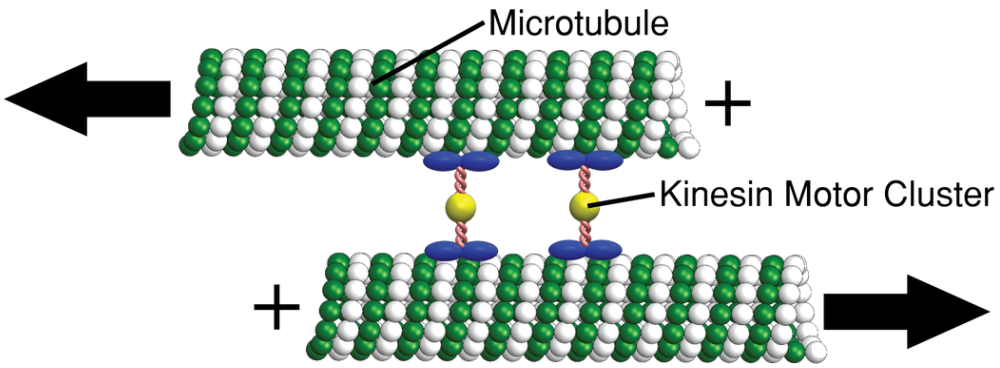
Macromolecules

Biological structure is critical to:

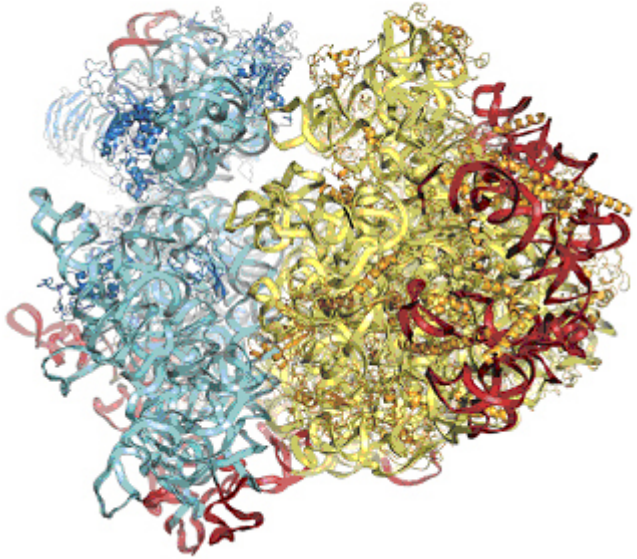
– Understanding how biology works
– Diagnosing, preventing, and treating disease
– Food and energy production (e.g., agriculture)

Structure key to function

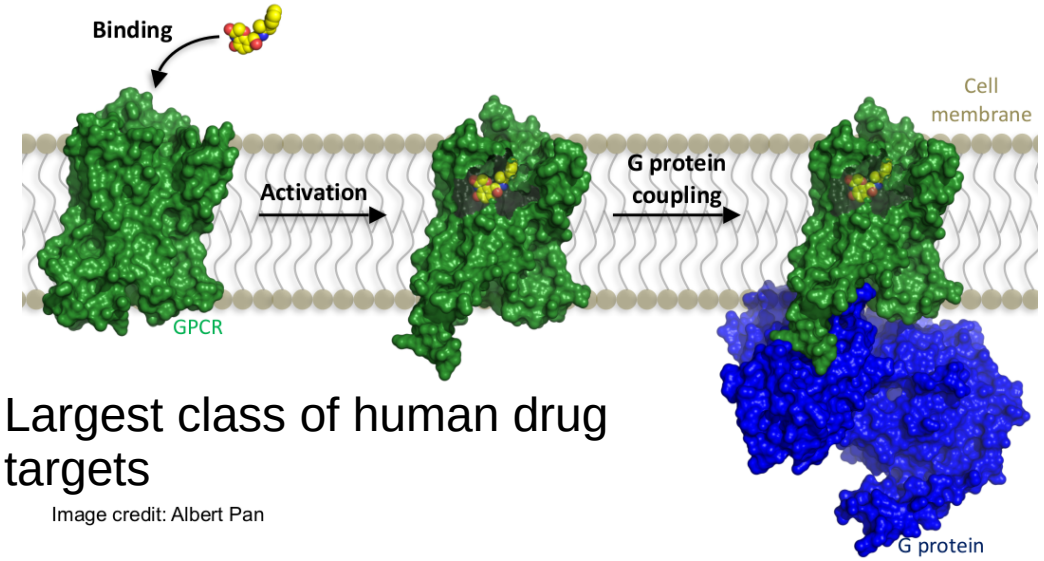
Kinesin: motor proteins, microtubes



Ribosome: translation



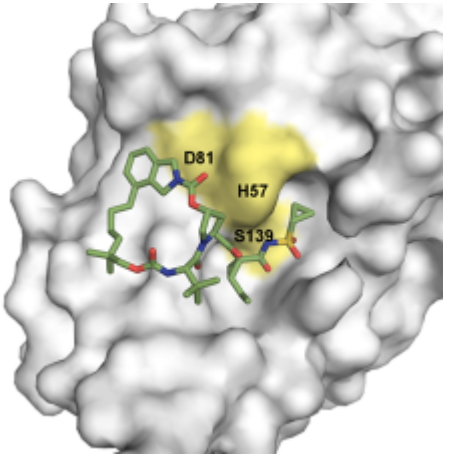
G-protein coupled receptors (GPCRs)



Largest class of human drug targets

Image credit: Albert Pan

Structure-based drug design



- Almost all drugs act by binding to proteins and altering their function
- Using knowledge of structures, we can
  - design drugs that bind tightly to the desired protein,
  - alter behavior of the protein in a desired way, avoid binding to other proteins, etc.

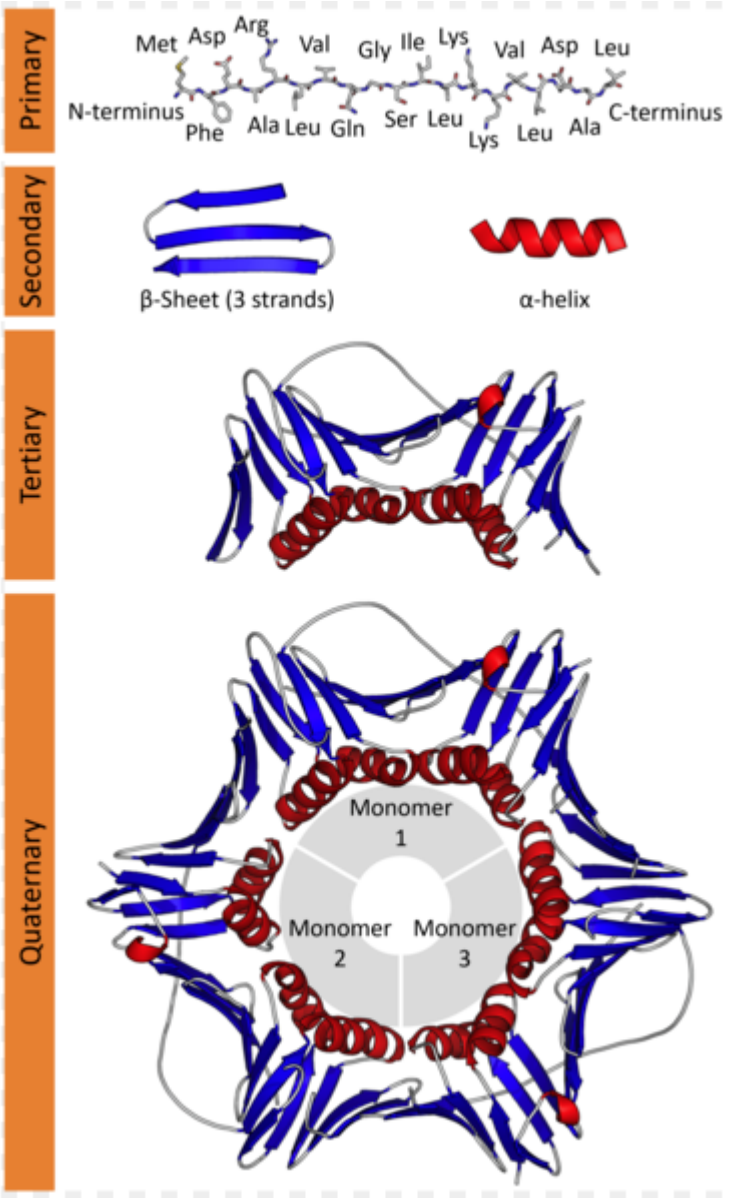
Protein structures

Primary structure
- sequence of amino acids in the polypeptide chain

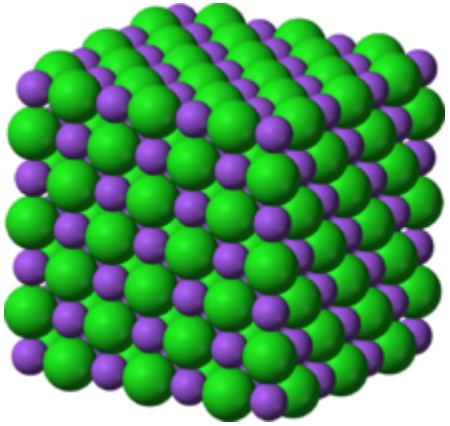
Secondary structure
- local sub-structures
- α-helix and β-sheets (two main types)

Tertiary structure
- the three-dimensional structure created by a single protein

Quaternary structure
- three-dimensional structure of all polypeptide chains (subunits) that operate as a single functional unit



Experimentally Determined Protein Structures

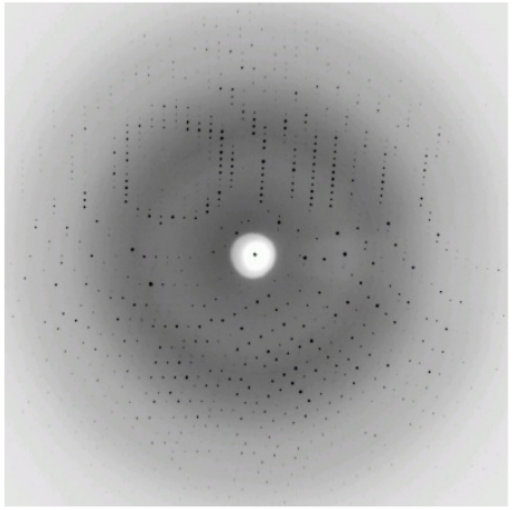


Crystallized Salt

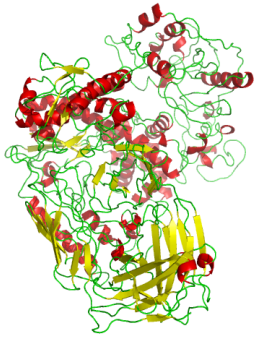
X-ray crystallography is the primary method for determining protein structure

- pure crystal of high regularity
- X-ray diffraction used to determine structure

But.. it is extremely difficult to predict good conditions for nucleation or growth of well-ordered crystals



X-ray diffraction pattern



Protein crystal structures: The Protein Data Bank (PDB)

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | Documentation | Careers | MyPDB

RCSB PROTEIN DATA BANK

189185 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

PDB Archive | Advanced Search | Browse Annotations | Help

PDB-101 | wwPDB | EMDataResource | NDB Nucleic Acid Database | Worldwide Protein Data Bank Foundation

Developers: Join the RCSB PDB Team — Explore Open Positions

- Welcome
- Deposit
- Search
- Visualize
- Analyze
- Download
- Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

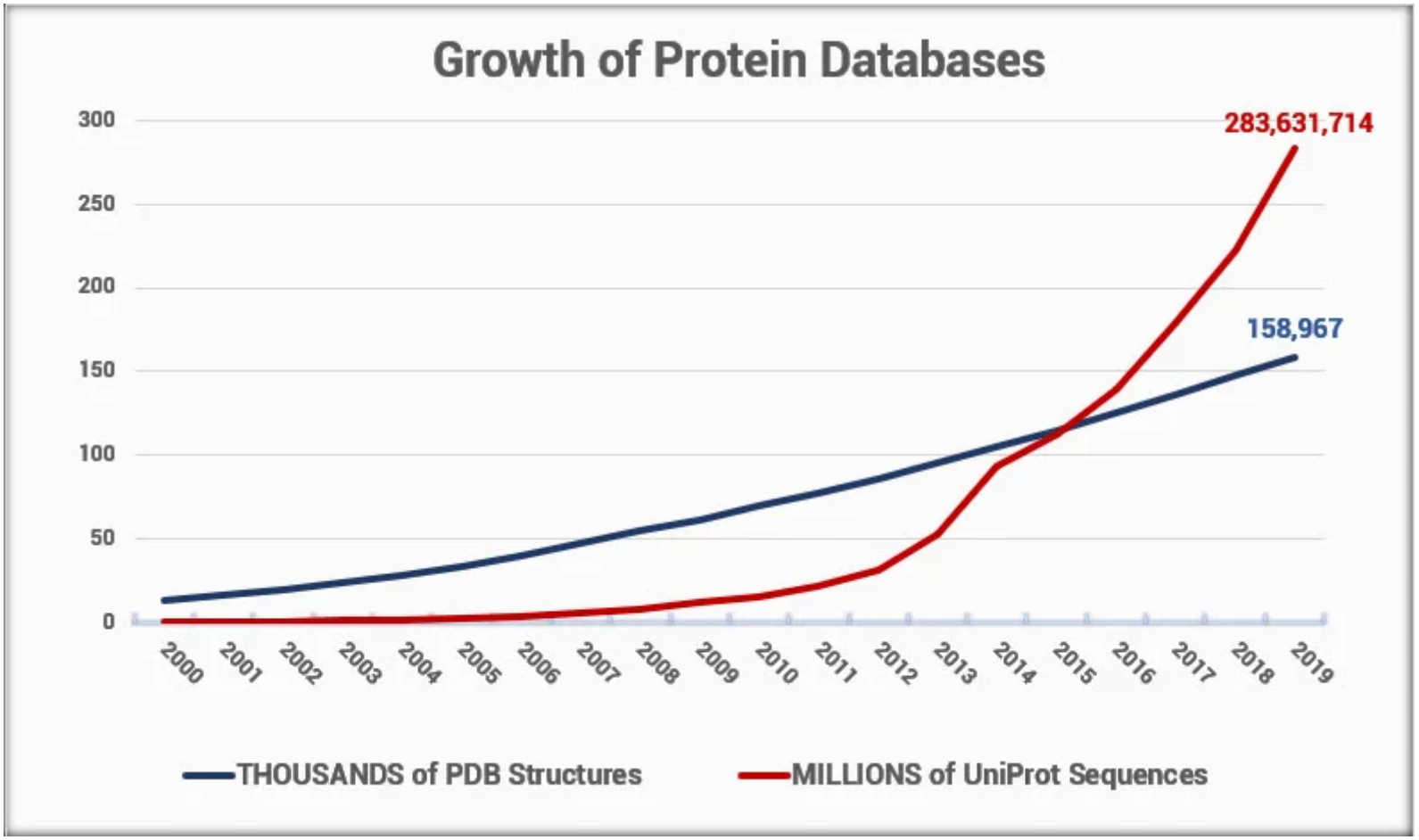
The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

April **Molecule of the Month**

- A collection of essentially all published, experimentally determined structures of biomacromolecules (e.g., proteins, DNA, RNA)
- Each identified by 4-character code (e.g., 6YYT)
- Currently ~189,000 structures. ~80% of those are determined by x-ray crystallography

HER2/neu and Trastuzumab

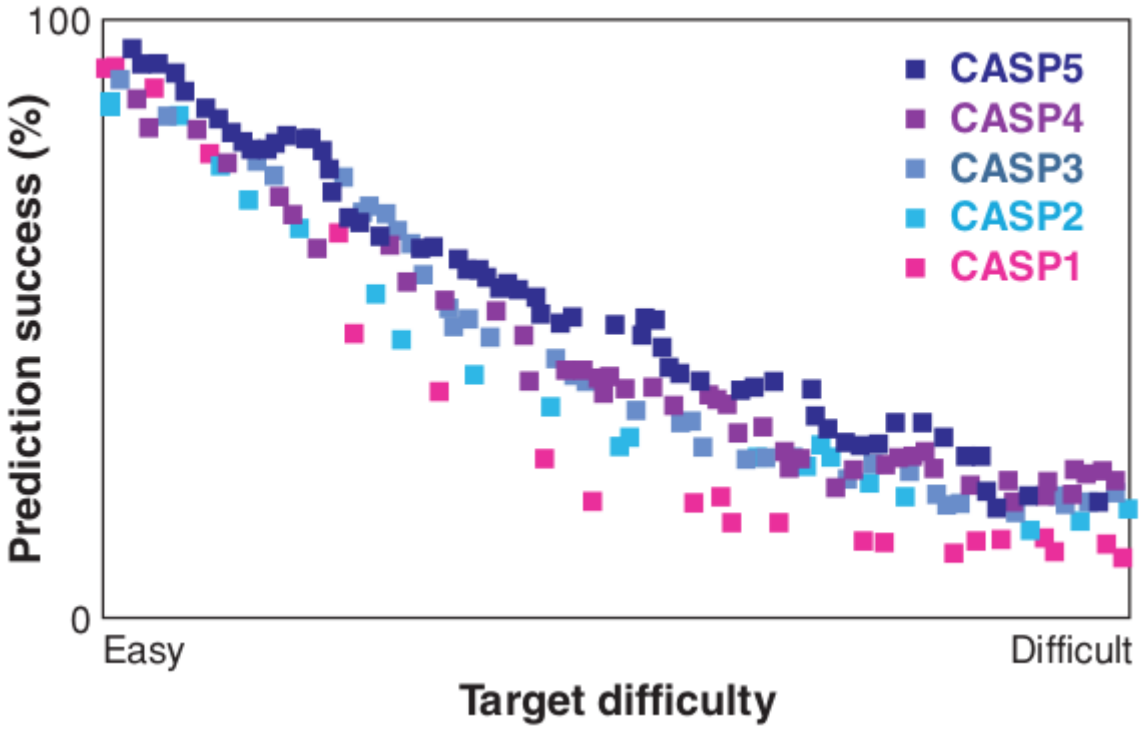
Most proteins have no known structure

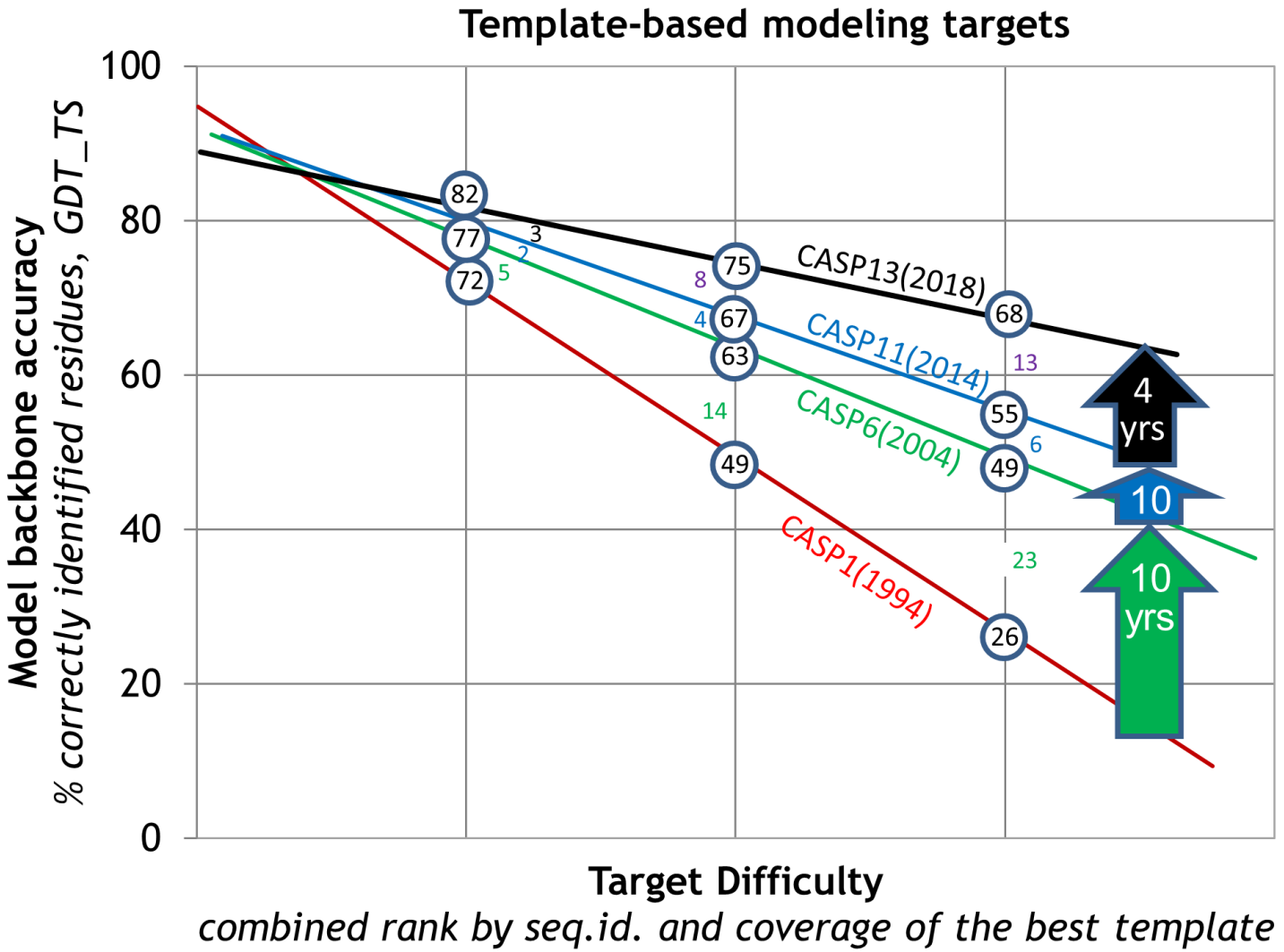


Protein Structure Prediction

Critical Assessment of protein Structure Prediction (CASP)

- A community competition for protein structure prediction
- Double blind experiment based on solved or soon to solve structures
- Both template (homology) and de novo predictions
- CASP1 (1994), competition is every two years





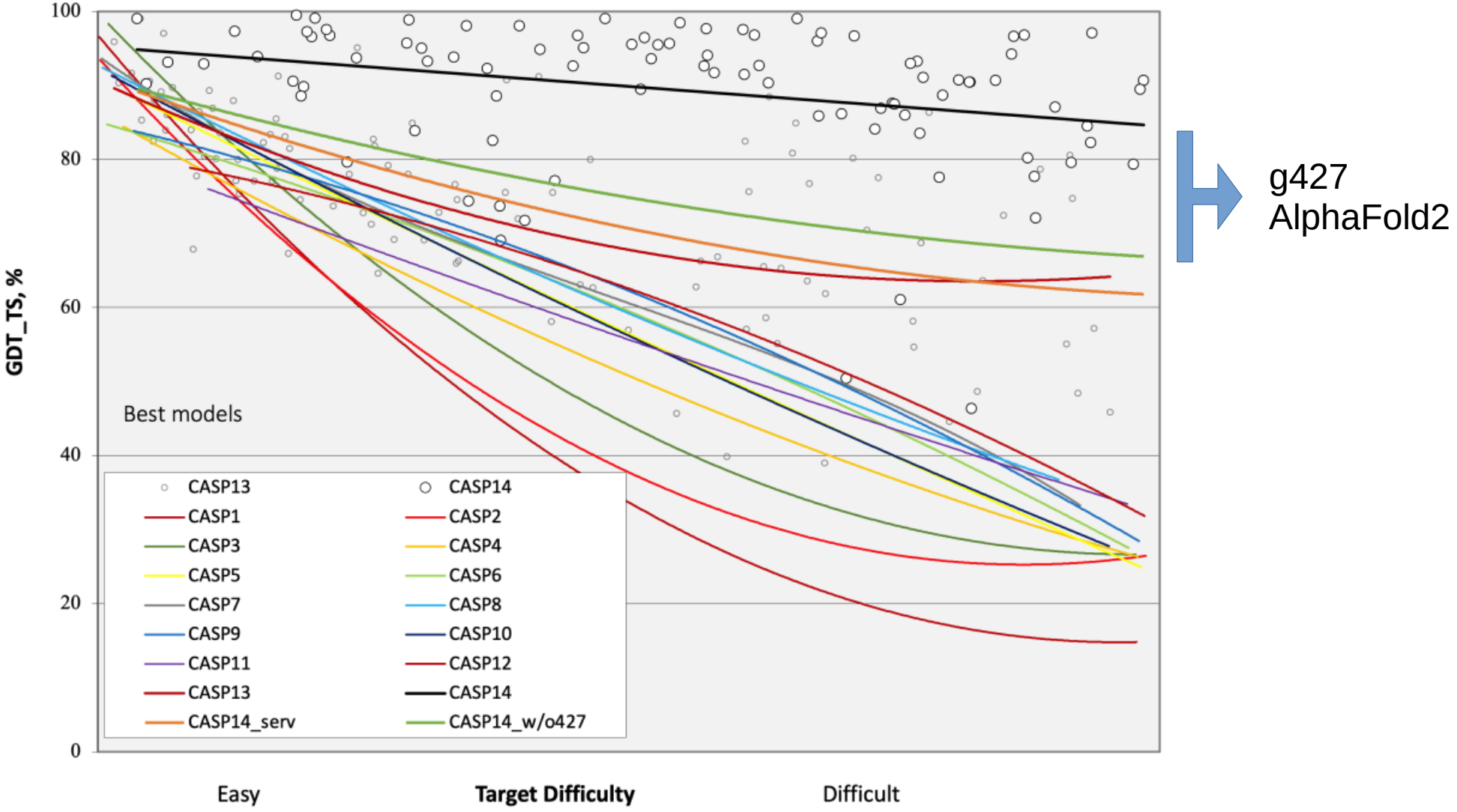
A burst in progress

GDT: global distance test, % of amino acids at correct position, 90% is competitive with experimental structures

Progress has been achieved by:

- more accurate alignment of the target sequence to that of available templates,
- combining multiple templates, improved accuracy of regions not covered by templates,
- successful refinement of models, and better selection of models from decoy sets due to improved methods for estimation of model accuracy.

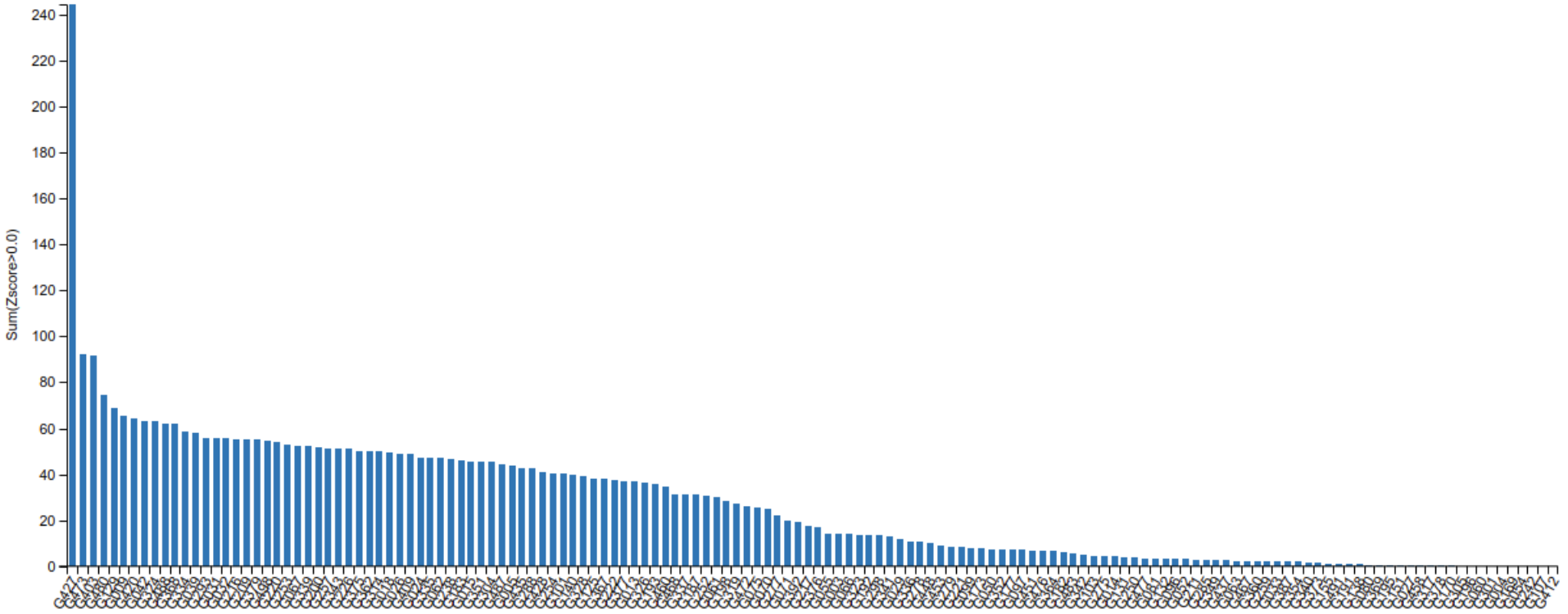
CASP14 2020



g427
AlphaFold2

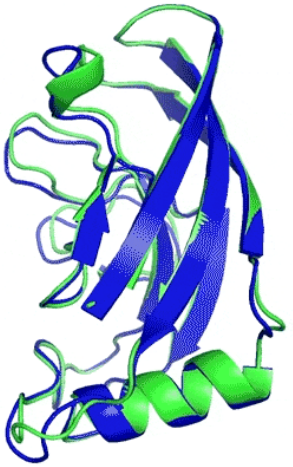
GDT: global distance test, % of amino acids at correct position, 90% is competitive with experimental structures

CASP14 participant scores



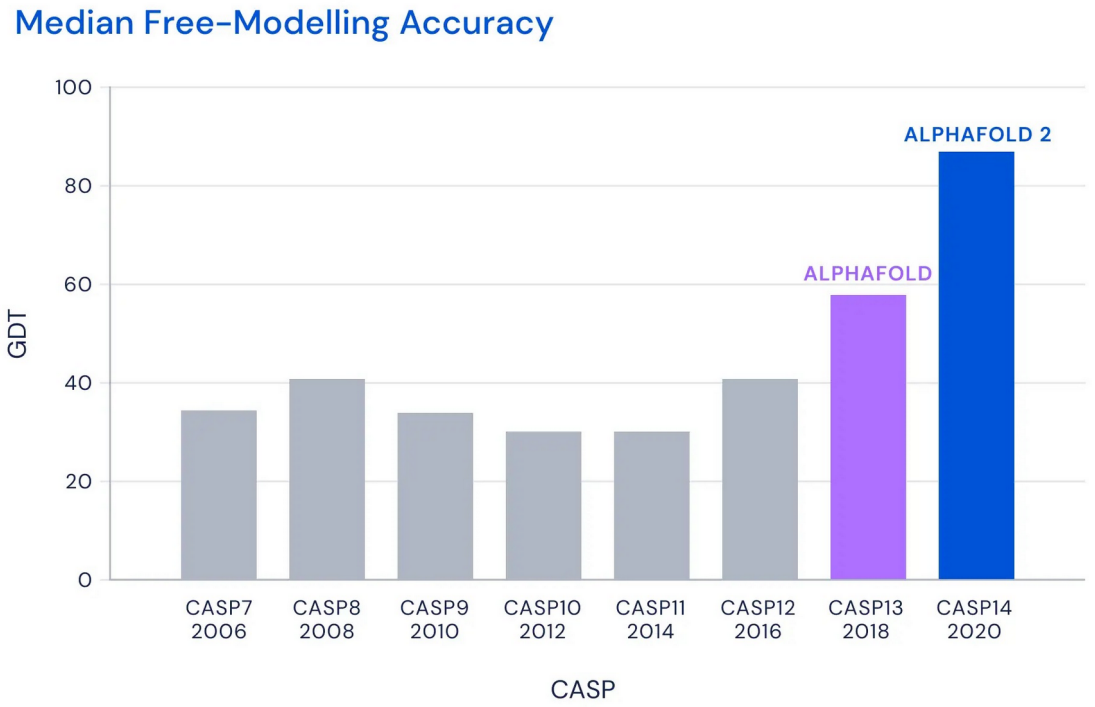
Predictions are often very accurate

T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)



T1049 / 6y4f
93.3 GDT
(adhesin tip)

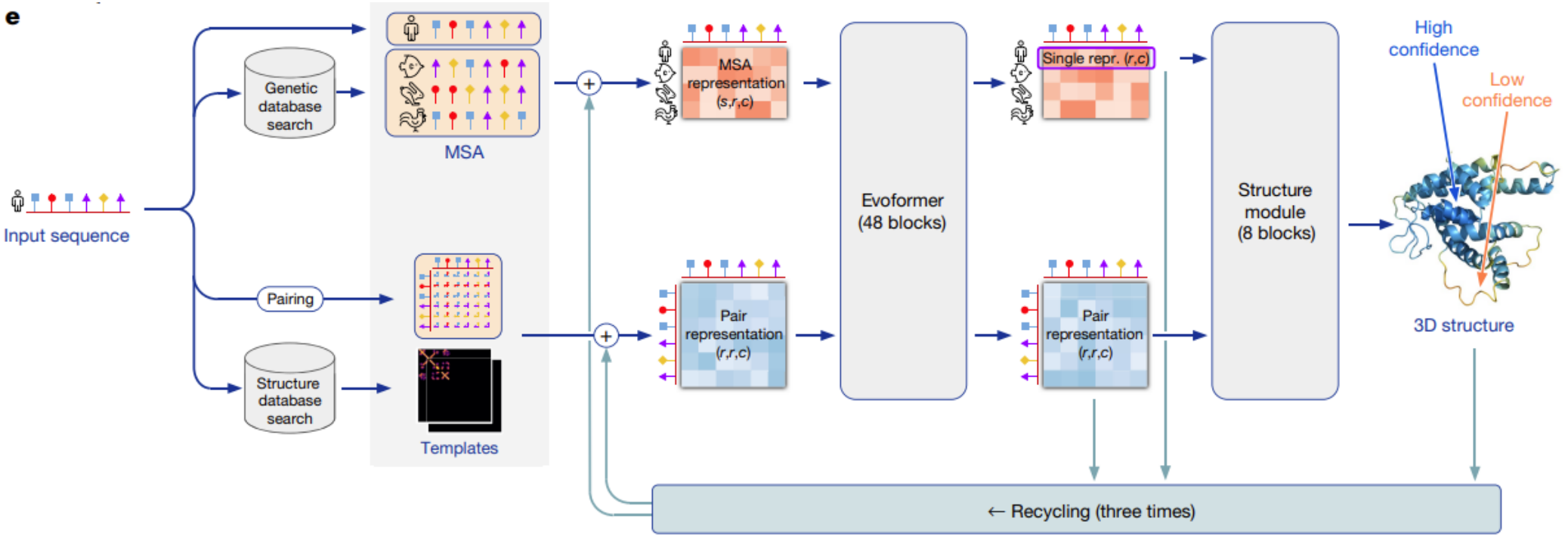
Experimental result
Computational prediction



AlphaFold2 (2021)

A Deep neural network:
- physical/statistical interactions
- evolution: constraints, homology models, pairwise correlations
- growth of PDB structure database

How does one predict 3D structure?



The Protein Folding Problem

Really three questions (grand challenges):

1) the folding code: the thermodynamic question of what balance of interatomic forces dictates the structure of the protein, for a given amino acid sequence

2) protein structure prediction: the computational problem of how to predict a protein's native structure from its amino acid sequence; and

3) the folding process: the kinetics question of what routes or pathways some proteins use to fold so quickly

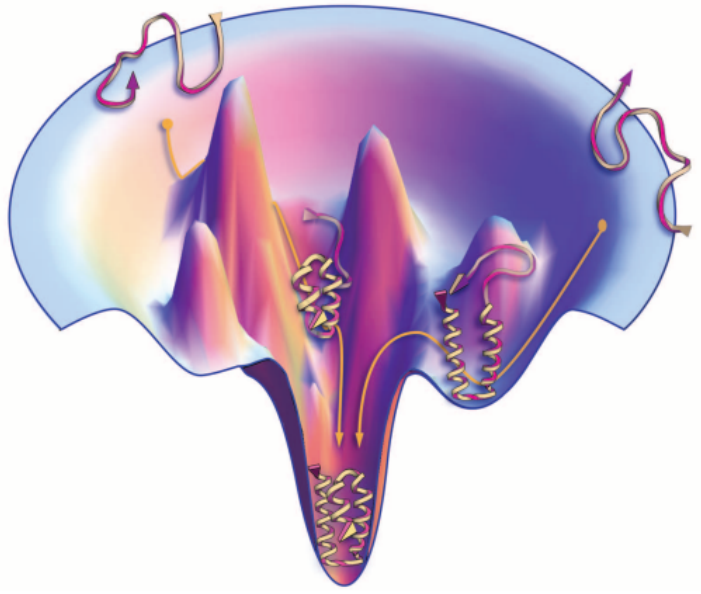
The folding code: Anfinsen's Thermodynamic Hypothesis

Hypothesis: the native structure is determined only by the protein's amino acid sequence, which determines a unique, stable and kinetically accessible minimum of the free energy

Experiment: Anfinsen unfolded the RNase enzyme under extreme chemical conditions and observed that the enzyme's amino acid structure refolded spontaneously back into its original form when he returned the chemical environment to natural cellular conditions

- in vitro studies of protein structure
- 3D structures can be predicted from physical chemistry

  - electrostatics: hydrogen bonds, ion pairs
  - van der Waals attractions
  - water-mediated hydrophobic interactions

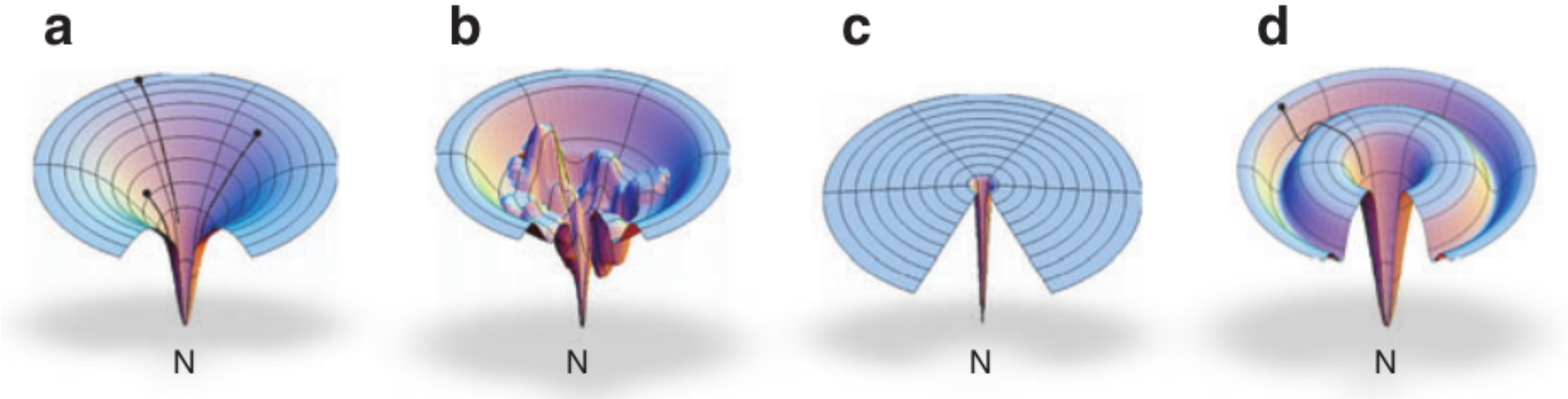


Energy landscapes

Intramolecular-plus-solvation free energy

Astronomical number of possible conformations to search

Funnel shape: many high and few low energy states (native)



Fast fold

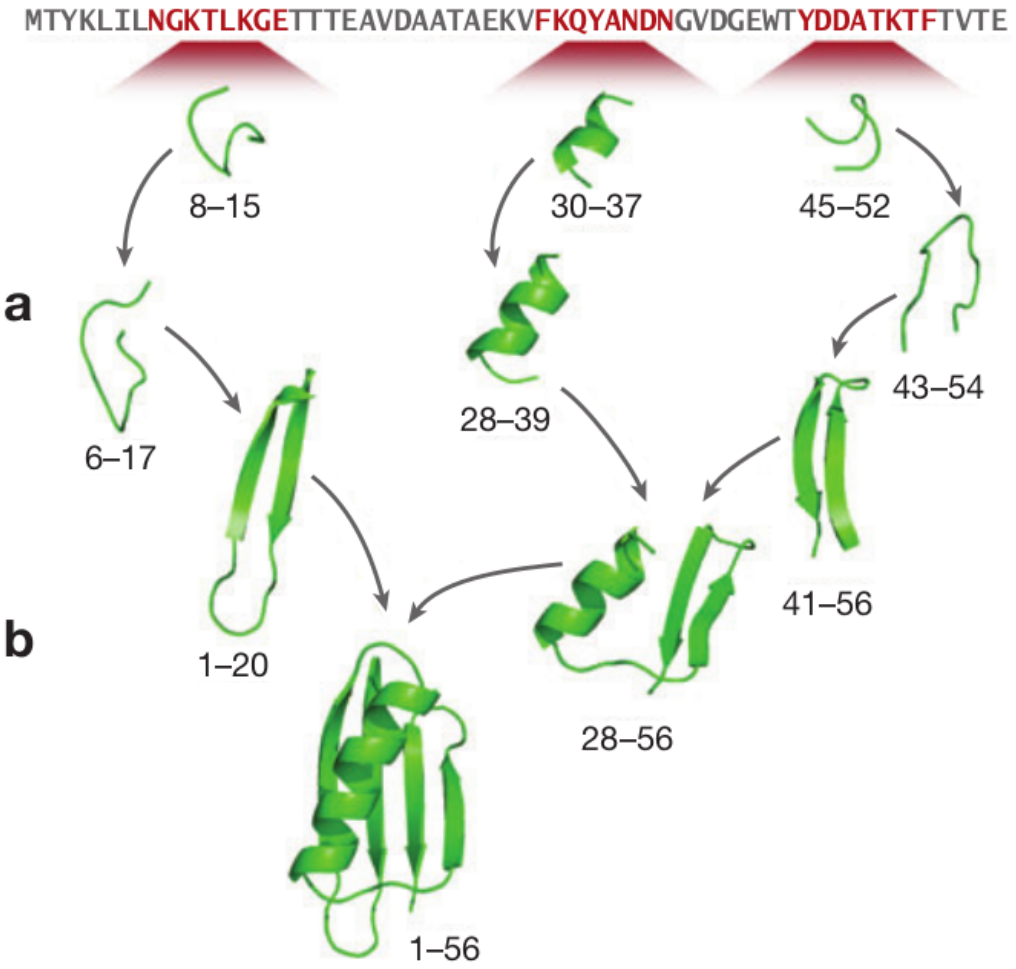
Rugged with kinetic trapping

Slow random

Moat with intermediate

Zipping and assembly

- small fragments of the chain can search their conformations more completely than larger fragments
- divide-and-conquer strategy
- peptide fragments first find local metastable structures, such as helical turns, β-turns, or small loops (in parallel)
- local structures are sufficiently metastable to enable larger stable structures to form



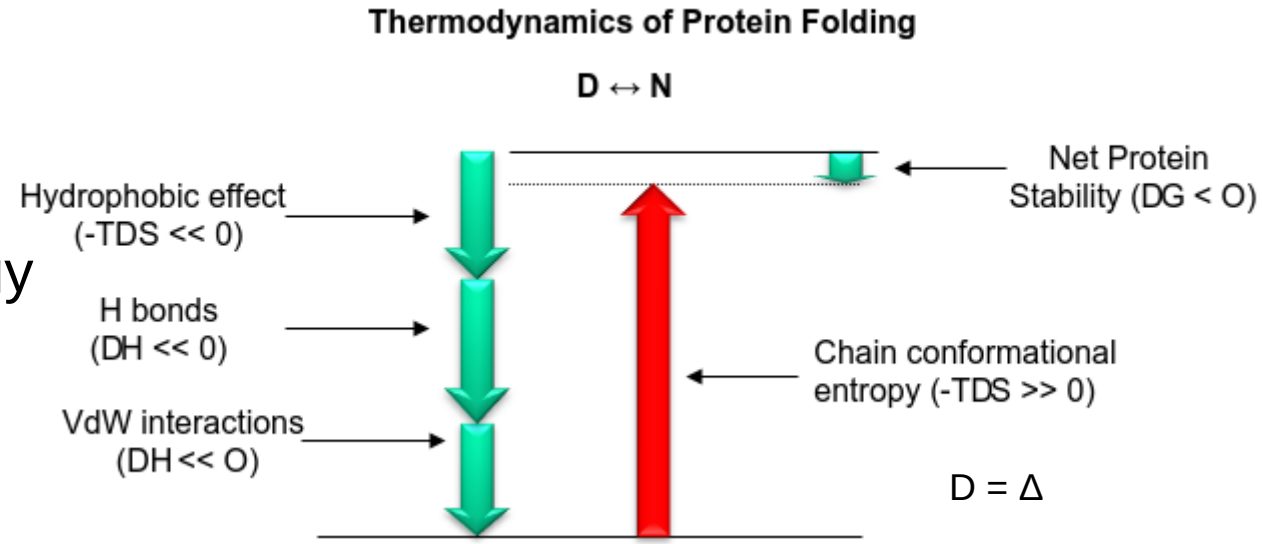
Protein stability

Thermodynamic stability of proteins represents the free energy difference between the folded and unfolded protein states

Gibbs free energy: the maximum reversible work that may be performed by a thermodynamic system

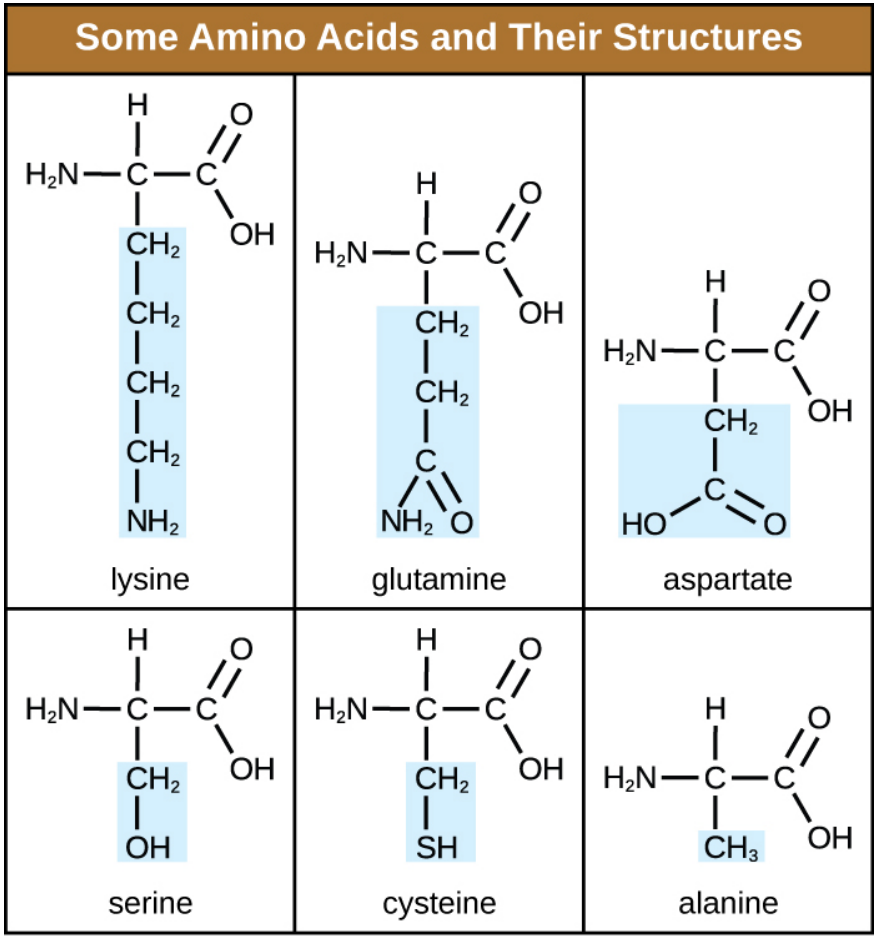
$\Delta G = \Delta H - T\Delta S$

ΔG depends on:
H enthalpy, internal energy
T temperature
S entropy, randomness



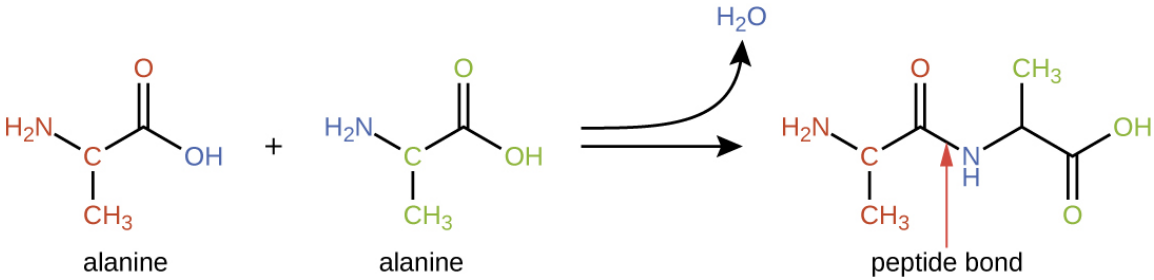
A negative change in free energy is favorable
Typical protein: ΔG between folded and unfolded is small 5-10 kcal/mol

Proteins: Review



*Blue shading indicates R group.

- Protein - a peptide chain, made of amino acids
- Amino acids – hydrogen, carboxyl group and amino group with a side chain (R)
- Amino acid are covalently bonded to form peptide chains

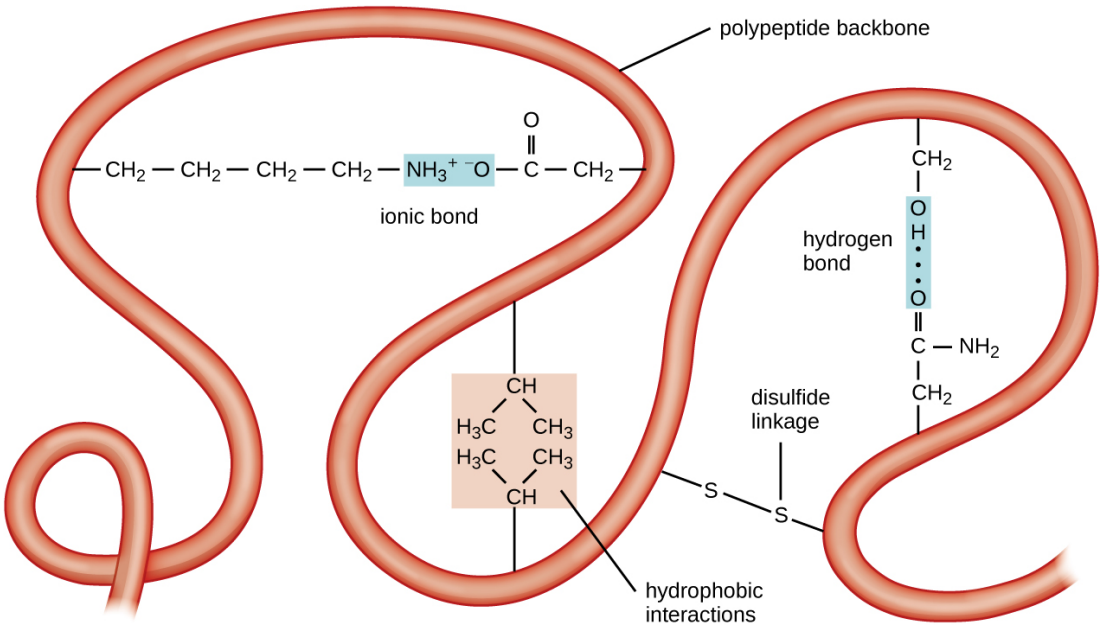


Determinants of protein stability

Non-covalent forces (covalent = sharing electrons)

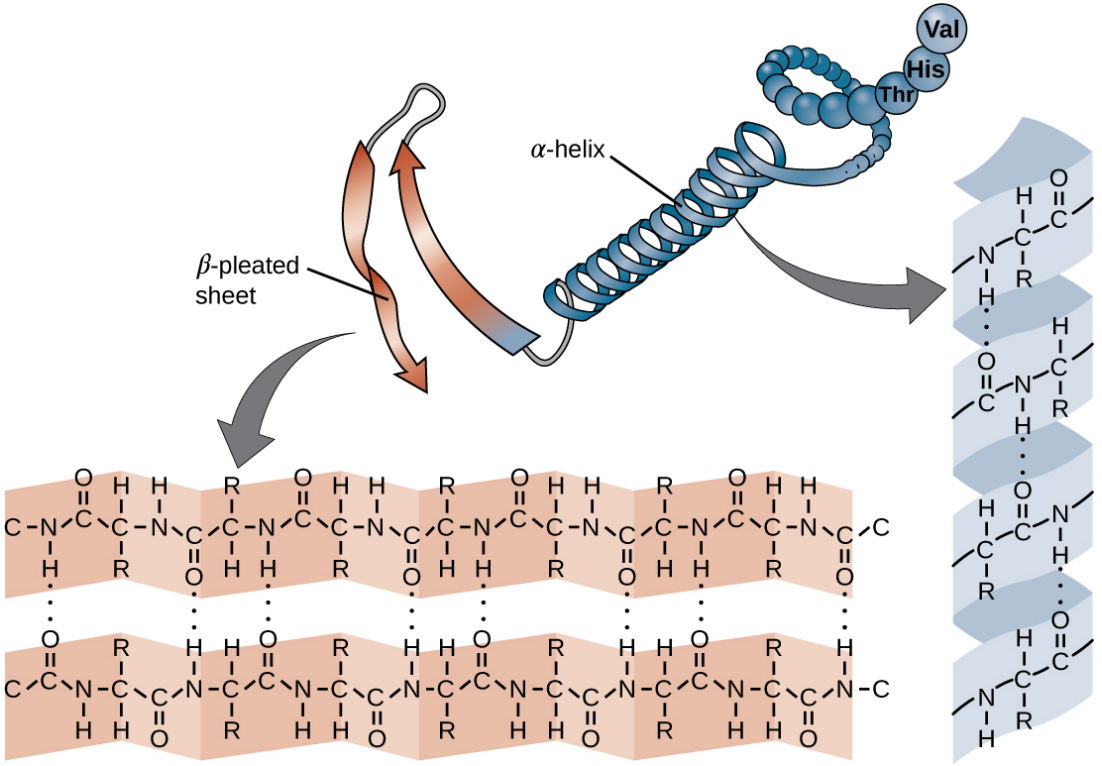
- native proteins are only 5–10 kcal/mol more stable than their denatured states, small intermolecular forces must be accounted for in folding and structure prediction

- Electrostatics (between electric charged particles, inverse square dist.)
  - Ion - Ion Interactions (e.g. oppositely charged amino acids)
    - but few ion pairs, and not conserved in evolution
  - Hydrogen bonding (e.g. with water or sidechains)
    - alpha helix, beta sheet



Hydrogen bonds: secondary structure

- Electrostatics (between electric charged particles, inverse square dist.)
  Ion - Ion Interactions
  Hydrogen bonding
  - but H bonds with water in the unfolded state
  - enthalpy (bonds) vs entropy (disorder), $\Delta G=\Delta H-T\Delta S$
  - H bonds don't drive folding but form so that the folded protein would not be destabilized by too many unsatisfied H bonds

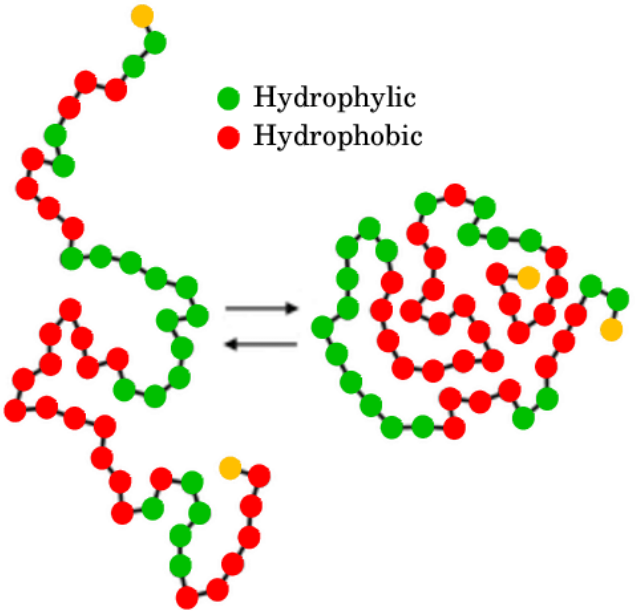


Determinants of protein stability

- Electrostatics (between electric charged particles, inverse square dist.)
  - Ion - Ion Interactions
  - Hydrogen bonding

- van der Waals (weak) forces
  - between fixed or induced dipoles
  - between hydrophobic (non-polar) side-chains
  - important since there is tight packing of folded proteins

- Hydrophobic interactions (dominant force)
  - most non-polar side chains are buried, hydrophobic core
  - proteins are denatured in non-polar solvents
  - replacing hydrophobic residues with polar ones destabilizes proteins
  - but.. can't be the only force

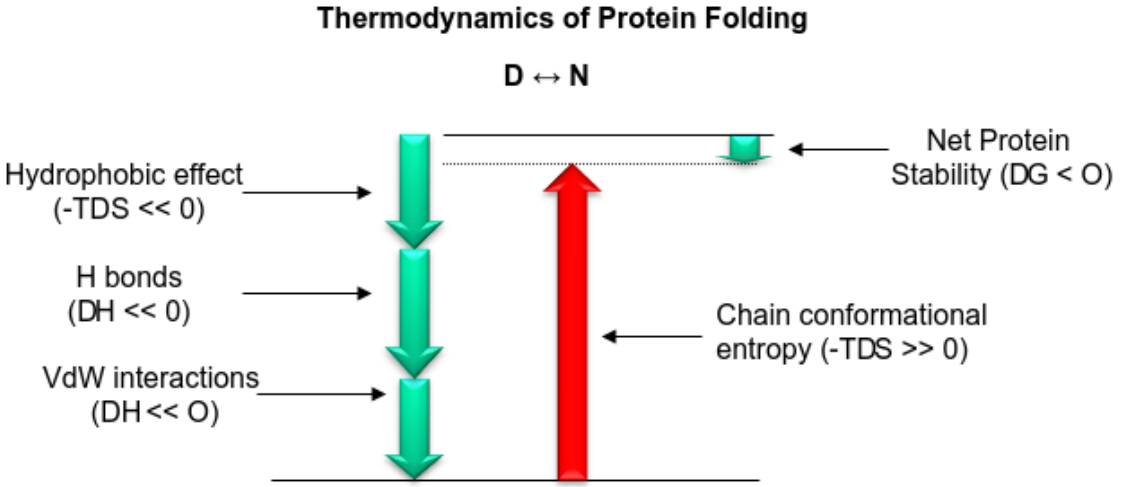


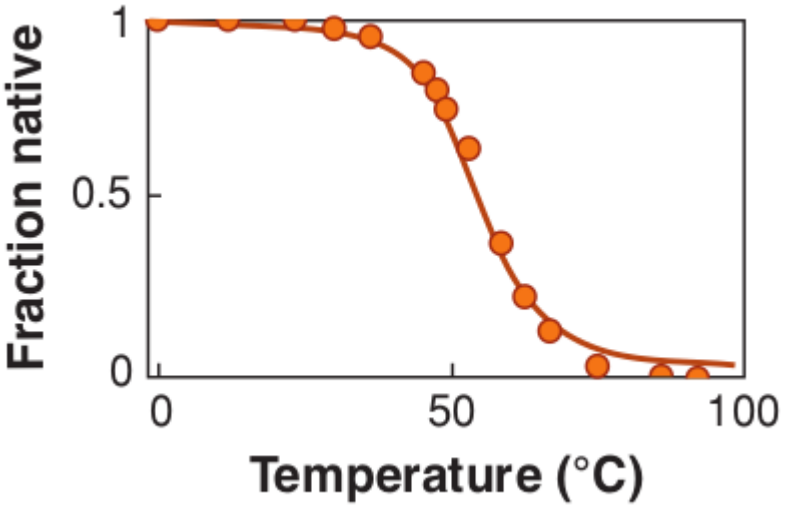
Free energy of folding

$\Delta G = \Delta H - T\Delta S$

Thermophilic organisms

- >50oC for optimal growth
- salt bridges increase but not H-bonds
- very compact (maximizing van der Waals interactions)

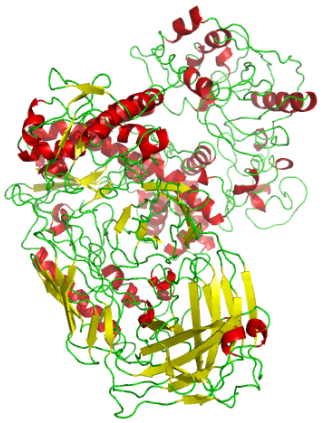




Protein structure prediction

Given the sequence of amino acids that make up a protein, predict its 3D structure

MWRLRSIARANIHICNQ
FLPVSNKSIGTLSVFRF
YSSSLEERYKEKLLQEA



Given the structure, what is the potential energy or enthalpy?

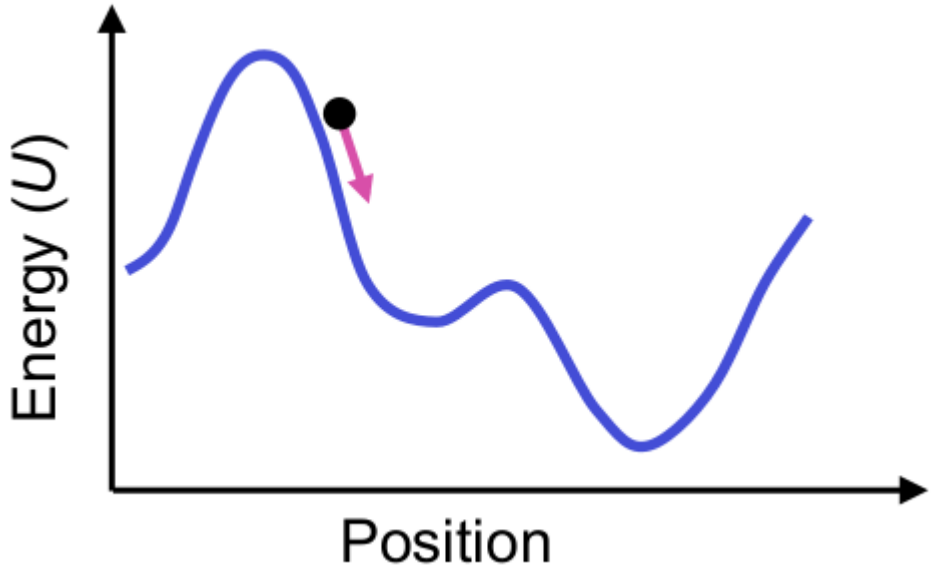
Potential energy is the energy of a system at any state, whereas free energy is the thermodynamic potential between two equilibrium states (enthalpy and entropy). We can approximate enthalpy.

Approximate Energy Functions

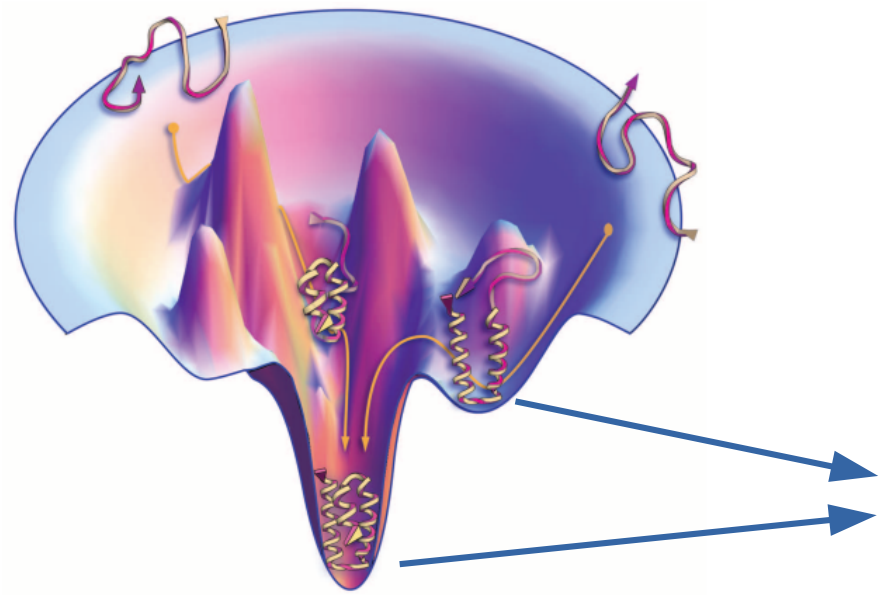
Energy and force

A potential energy function U(x) specifies the total potential energy of a system of atoms as a function of all their positions (x)

- Force on atom i is given by derivatives of U with respect to the atom's coordinates x_i , y_i , and z_i $F(x) = -\nabla U(x)$
- At local minima of the energy U, all forces are zero
- The potential energy function U is also called a force field



Energy functions



Potential (quasi-chemical approximation) energy

A wide variety of force fields are used in atomic-level modeling of macromolecules

• Physics-based vs. knowledge-based

– Physics-based force fields attempt to model actual physical forces

– Knowledge-based force fields are based on statistics about, for example, known protein structures

– Most real force fields are somewhere in between

• Atoms represented

– Most realistic choice is to model all atoms

– Some force fields omit waters and other surrounding molecules. Some omit certain atoms within the protein.

Molecular mechanics force fields

Molecular mechanics force

- used for molecular dynamics simulations
- more toward the physics-based, all-atom end (i.e., the more “realistic” force fields)
  - – Represent physical forces explicitly
  - – Typically represent solvent molecules (e.g., water) explicitly
- Forces:
  - Bond length stretching
  - Bond angle bending
  - Torsional angle twisting
  - Electrostatics interaction
  - van der Waals interaction

Exercises

What determines protein stability?

List two types of secondary structure and what type of interactions mediate those structures?

How do we know a protein's structure is determined by amino acid sequence?

Solvant (e.g. water) are not important to protein structure/stability [T/F]

What is force plays a dominant role in protein folding?

Entropy important in protein folding [T/F]?