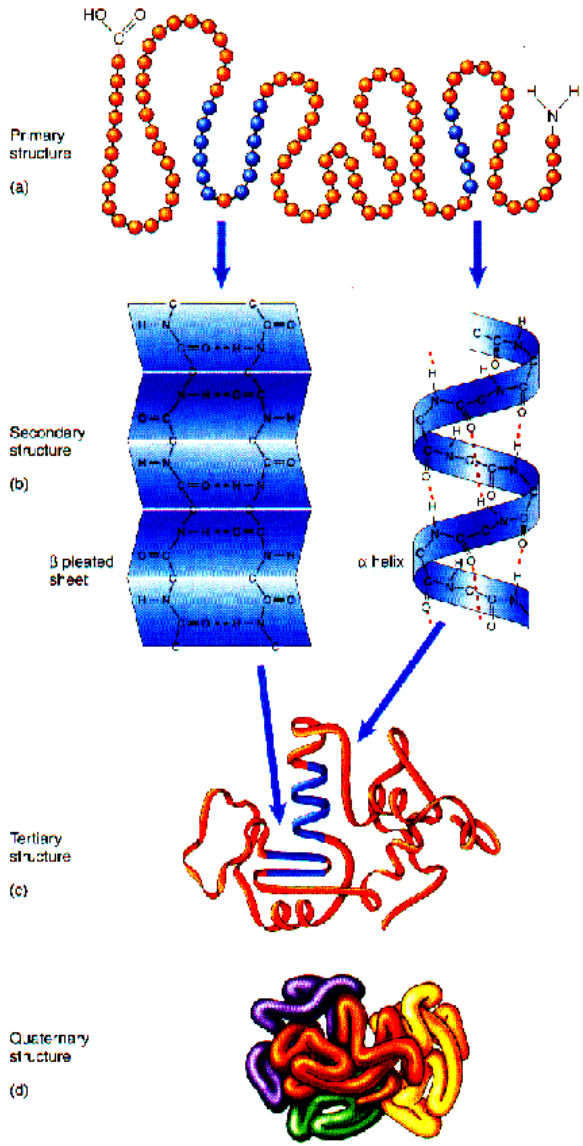
CS612 - Algorithms in Bioinformatics

Protein Folding

March 1, 2023

The Protein Folding Problem

- Protein folding is the translation of primary sequence information into secondary, tertiary and quaternary structural information
- Don't forget post-translational modifications.
- They change the chemical nature of the primary sequence and thus affect the final structure



Relationship Between Structure and Function

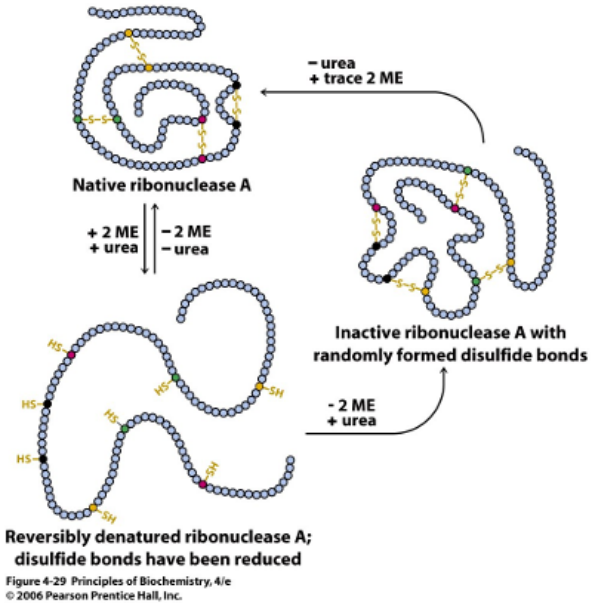
- H. Wu, 1931: First formulation of the protein folding problem on record
- Mirsky and Pauling 1936:
  - Chemical and physical properties of protein molecules attributed to amino-acid composition and structural arrangement of the amino-acid chain
  - Denaturing conditions like heating assumed to abolish chemical and physical properties of a protein by "melting away" the protein structure
- The relationship between protein structure and function was under significant debate until the revolutionizing experiments of Christian Anfinsen and colleagues at the National Institute of Health (NIH) in the 1960s

Anfinsen's Experiments: Spontaneous Refolding

- Experiments by Christian B. Anfinsen showed that the small ribonuclease enzyme would re-assume structure and enzymatic activity after denaturation
- This ability to regain both structure and function was confirmed on thousands of other proteins.
- It seemed to be an inherent property of amino-acid chains
- After a decade of experiments, Anfinsen concluded that the amino-acid sequence governed the folding of a protein chain into a "biologically-active conformation" under a "normal physiological milieu"
- He received the Nobel prize in chemistry for his formulation of this relationship between the amino acid sequence and the biologically-active (functional) structure of a protein

Christian Anfinsen
The Nobel Prize in Chemistry 1972

The telegram that I received from the Swedish Royal Academy of Sciences specifically cites "...studies on ribonuclease, in particular on the relationship between the amino acid sequence and the biologically active conformation..."
"Studies on the Principles that Govern the Folding of Protein Chains"
Nobel Lecture, Dec. 11, 1972



From Anfinsen's Experiments to Simulations

- Let the 3D spatial arrangement of atoms constituting the protein chain be referred to as a conformation
- How does the amino-acid sequence determine the biologically-active conformation?
- If one knows the answer to the above, one can formulate instructions for a computer algorithm to compute the biologically-active conformation
- There are many conformations (possible arrangements of the chain) of a protein chain
- What makes the biologically-active conformation different from the rest?
- Can this information be used to "guide" a computer algorithm to this special conformation?

Christian Anfinsen
The Nobel Prize in Chemistry 1972

The telegram that I received from the Swedish Royal Academy of Sciences specifically cites "...studies on ribonuclease, in particular on the relationship between the amino acid sequence and the biologically active conformation..."
"Studies on the Principles that Govern the Folding of Protein Chains"
Nobel Lecture, Dec. 11, 1972

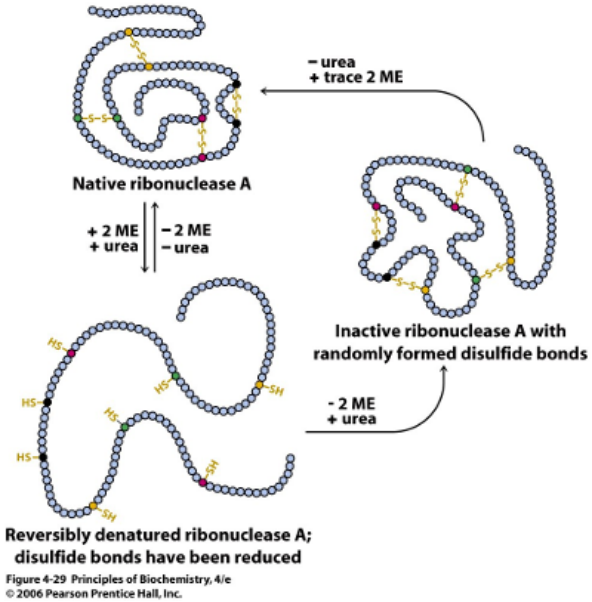


Figure 4-29 Principles of Biochemistry, 4/e
© 2006 Pearson Prentice Hall, Inc.

The Thermodynamic Hypothesis

Anfinsen' experiments made the case for the thermodynamic hypothesis:

- This hypothesis states that the three-dimensional structure of a native protein in its normal physiological milieu (solvent, pH, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, etc.) is the one in which the Gibbs free energy of the whole system is lowest.
- That is, that the native conformation is determined by the totality of inter-atomic interactions and hence by the amino acid sequence, in a given environment.

[Anfinsen's Nobel Lecture, December 11, 1972]

Case Study... What Happens When it All Goes Wrong

Eligibility criteria for blood donation, from the red cross website:

"You are not eligible to donate if:

From January 1, 1980, through December 31, 1996, you spent (visited or lived) a cumulative time of 3 months or more, in the United Kingdom (UK), or From January 1, 1980, to present, you had a blood transfusion in any country(ies) in the (UK) or France."

Case Study... What Happens When it All Goes Wrong

- A bovine epidemic struck the UK in 1986, 170,000 cows appeared to be mad: they drooled and staggered, were extremely nervous, or bizarrely aggressive. They all died. As the brains of the dead "mad" cows resembled a sponge, the disease was called bovine spongiform encephalopathy, or BSE.
- Other examples of spongiform encephalopathy are scrapie which develops in sheep, Creutzfeld-Jacob Disease (CJD) and its variant (vCJD) which develop in humans.

- In 1982 the infectious agents responsible for transmitting spongiform encephalopathy were defined and named prions (Stanley Prusiner, 1982).

The Mad Cow... What Happens When it All Goes Wrong

- Prions are proteins that are found in the nerve cells of all mammals. Many abnormally-shaped prions are found in the brains of BSE-infected cows and vCJD or CJD patients.
- The difference in normal and infectious prions may lie in the way they fold.
- Evidence indicates that the infectious agent in transmissible spongiform encephalopathy is a protein.
- The normal protein is called PrPC (for cellular). Its secondary structure is dominated by alpha helices.

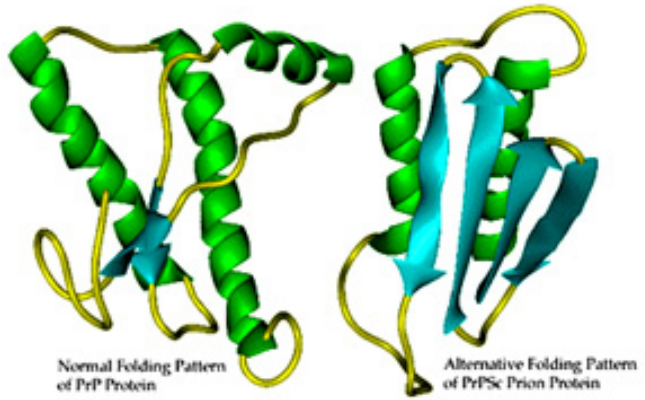
- The abnormal, disease producing protein called PrPSc (for scrapie), has the same primary structure as the normal protein, but its secondary structure is dominated by beta conformations.

The "Kiss of Death" or "Attack of the Zombies"

- The abnormally-shaped prion gets absorbed into the bloodstream and crosses into the nervous system.
- The abnormal prion touches a normal prion and changes the normal prion's shape into an abnormal one, thereby destroying the normal prion's original function.
- Both abnormal prions then contact and change the shapes of other normal prions in the nerve cell.
- The nerve cell tries to get rid of the abnormal prions by clumping them together in small sacs.
- Because the nerve cells cannot digest the abnormal prions, they accumulate in the sacs that grow and engorge the nerve cell, which eventually dies.
- When the cell dies, the abnormal prions are released to infect other cells. Large, sponge-like holes form where many cells die.

Prion Misfolding

- The Prion Hypothesis suggests that diseases like mad cow and human CJD are caused by the misfolding of a protein known as PrP that most cells contain.
- Once a few copies of the protein become misfolded), they cause other PrPs to misfold, leading to an accumulation of insoluble proteins in the cell.
- misfolded proteins cause cell death and damage the nervous system.



Not Only Cows: Human Prion Diseases

- Creutzfeldt-Jakob disease (CJD) is a rare fatal brain disorder that usually occurs in late life and runs a rapid course.
- No known treatment or cure.
- Most CJD cases are sporadic (not hereditary).
- It can also be acquired through contact with infected brain tissue (iatrogenic CJD) or consuming infected beef
- About 5 to 10% of cases are due to an inherited genetic mutation associated with CJD (familial CJD)
- The mutation makes the prion protein more susceptible to misfolding.
- How and why? Still not entirely clear...

Similar yet Different: Huntington's Disease

- A (usually) hereditary disease that causes death of brain cells
- Autosomal dominant inheritance pattern
- Symptoms: Mood disorders, uncoordinated movements, eventually dementia
- Typical age at onset: 30-50
- Life expectancy: 15-20 years from diagnosis

Woody Guthrie, 1912 – 1967

Dr. Remy Hadley ("13"), House MD

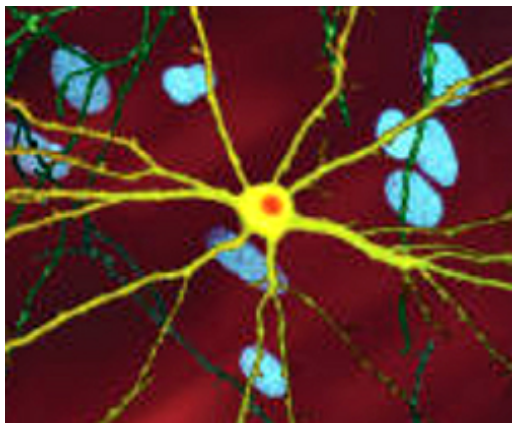
Trinucleotide Repeat Disorder

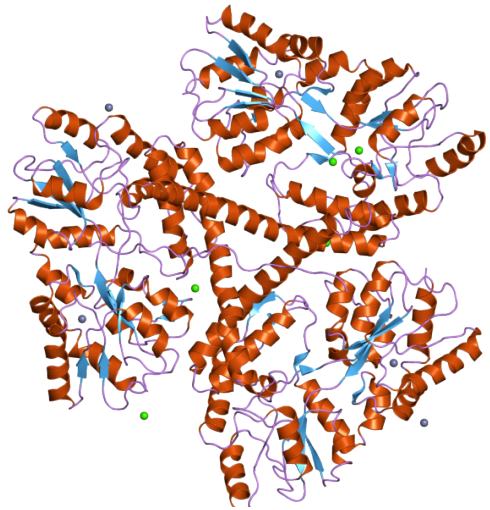
- The length of a repeated section of a gene exceeding a normal range
- The HTT gene contains a sequence of CAG– repeated multiple times (i.e. ... CAGCAGCAG ...)
- This sequence codes to Glutamine (GLN, Q)
- Number of repeats varies normally in the population
- This creates a sequence of Glutamines, called PolyQ (polyglutamine chain).
- An abnormally large number of CAG repeats in the HTT gene causes HD.

Repeat count	Classification	Disease status	Risk to offspring
<26	Normal	None	None
27–35	Intermediate	None	Elevated but << 50%
36–39	Reduced Penetrance	Maybe	50%
40+	Full Penetrance	Will be affected	50%

Trinucleotide Repeat Disorder

- A sequence of 36 or more glutamines results in the production of a protein which has different characteristics.
- The PolyQ regions appears to adopt a β-sheet structure
- This altered form, called mutant huntingtin (mHTT), increases the decay rate of certain types of neurons.
- The huntingtin protein interacts with over 100 other proteins, and appears to have multiple biological functions.
- Enzymes in the cell often cut the elongated protein into fragments, which form abnormal clumps inside nerve cells, and may attract other, normal proteins into the clumps (Sounds familiar?)





Computational Modeling of Protein Folding

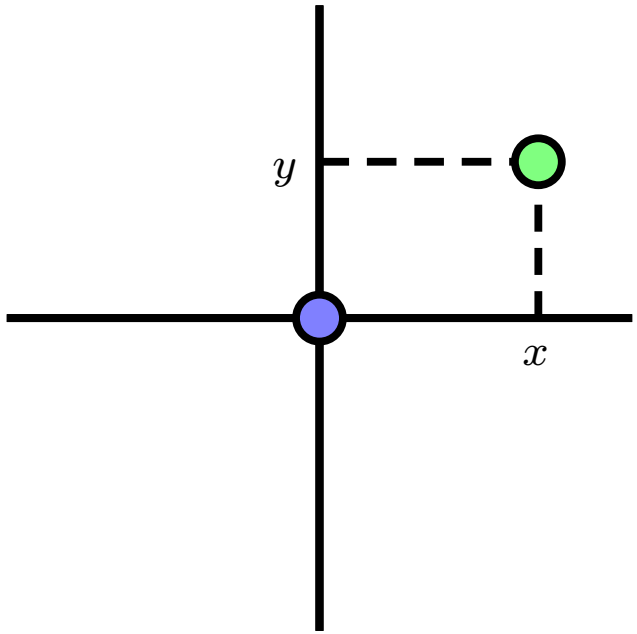
- Problem definition: Given the amino acid sequence, compute the correct 3-D arrangement (folded structure) of the protein.
- 3D coordinates representing locations of atoms.
- An energy function representing physical interactions.
- It should model how atoms interact with each other.
- Tells us how physically favorable is a given arrangement of atoms in space (conformation).
- We should able to compute it for every given conformation.
- Thermodynamics states that a molecule aims to fold into its minimum energy conformation.
- Exploration of the dynamics of proteins $\rightarrow$ search in the space spanned by the possible structures, guided by the energy function.

Degrees of Freedom (DOFs)

Definition (Degrees of Freedom)

The degree of freedom (DOF) is the set of independent parameters that can be varied to define the state of the system

Examples:
The location of a point in a 2-D cartesian system has two independent parameters – its (x, y) coordinates.



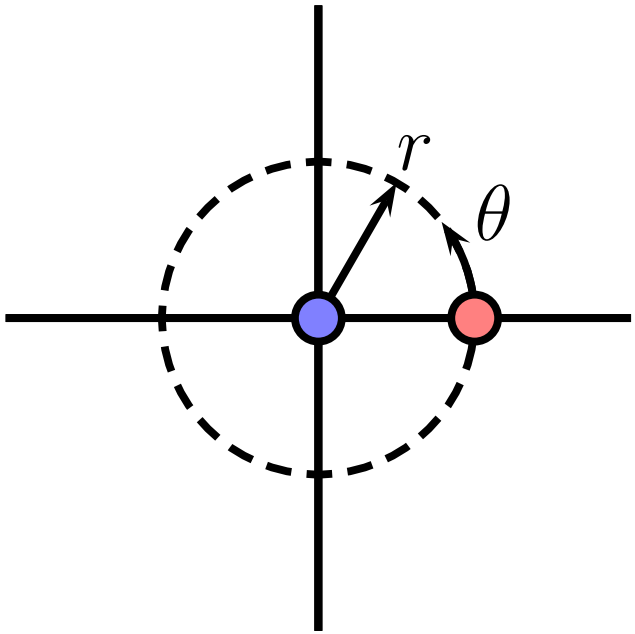
Degrees of Freedom (DOFs)

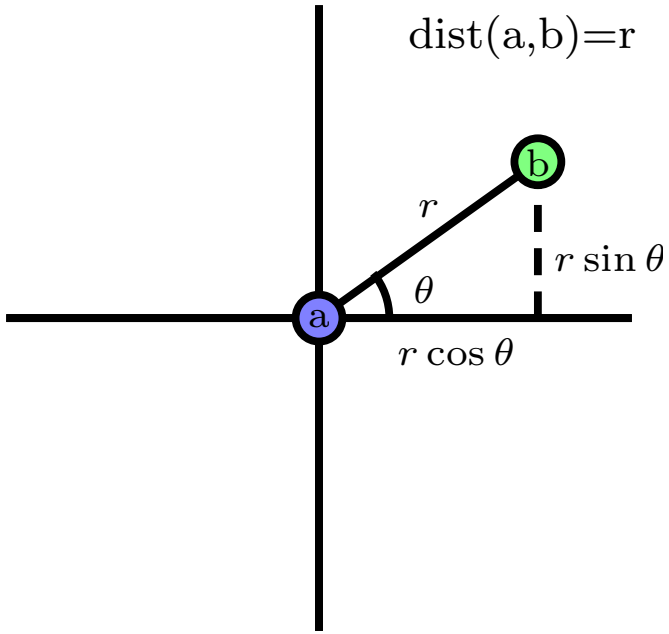
Definition (Degrees of Freedom)

The degree of freedom (DOF) is the set of independent parameters that can be varied to define the state of the system

Examples:

An alternative representation – (r, θ), distance from the origin and rotation about the origin, respectively.





Degrees of Freedom (DOFs)

Definition (Degrees of Freedom)

The degree of freedom (DOF) is the set of independent parameters that can be varied to define the state of the system

Examples:

- The location of a point in a 2-D cartesian system has three independent parameters – its (x, y) coordinates.
- An alternative representation – (r, θ), distance from the origin and rotation about the origin, respectively.
- A molecule with n atoms can be represented by a set of $3 \times N$ cartesian coordinates, so it has $3 \times N$ DOFs...
- Or does it?
- The actual number of DOFs is smaller, since distance and angle constraints restrict the atomic movement.

Representation by Internal Coordinates

- When trying to manipulate the structure internal coordinates may be easier to work with.
- The internal coordinates represent bond length, angles and dihedrals.
- Remember that we treat bond lengths and planar angles as fixed, but we still need them.
- They help us infer the connectivity of the structure and switch between representations.

Representation by Internal Coordinates

- Representing protein conformations with the dihedral angles as the only underlying degrees of freedom is known as the **idealized** or rigid geometry model.
- Ignoring bond lengths and bond angles greatly reduces the number of degrees of freedom and therefore the computational complexity of representing and manipulating protein structures.

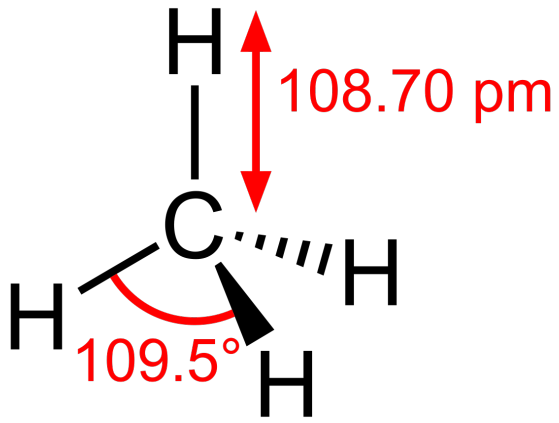
Representation by Internal Coordinates

- As a reminder – there are two freely rotatable backbone dihedral angles per amino acid residue in the protein chain: ϕ is a consequence of the rotation about the bond between N and Cα, and ψ , which is a consequence of the rotation about the bond between Cα and C.
- The peptide bond between C of one residue and N of the adjacent residue is not rotatable.
- The number of backbone dihedrals per amino acid is 2 (except the first and last), a total of 2N-2.
- but the number of side chain dihedrals varies with the length of the side chain. Its value ranges from 0, in the case of glycine, which has no side chain, to 5 in the case of arginine.

Representation by Internal Coordinates

- One can generate different three dimensional structures of the same protein by varying the dihedral angles.
- There are 2N-2 backbone dihedral DOFs for a protein with N amino acids, and up to 4N side chain dihedrals that one can vary to generate new protein conformations.
- Changes in backbone dihedral angles generally have a greater effect on the overall shape of the protein than changes in side chain dihedral angles (why?)

Methane Example



An example of a Z-matrix representing the internal coordinates of methane (CH4)

Atom	Bonded	Dist	Angle	Value	Dihe	Value
C						
H	1	1.089				
H	1	1.089	2	109.471		
H	1	1.089	2	109.471	3	120.0
H	1	1.089	2	109.471	3	-120.0

Methane Example

The Cartesian coordinate representation of Methane (CH4)

Atom	X	Y	Z
C	0.000	0.000	0.000
H	0.000	0.000	1.089
H	1.027	0.000	-0.363
H	-0.513	-0.889	-0.363
H	-0.513	0.889	-0.363

We can switch back and forth between different representations, up to an arbitrary rigid transformation (absolute position and orientation in space). To move from internal to cartesian coordinates we need the first three atoms, a, b, c.

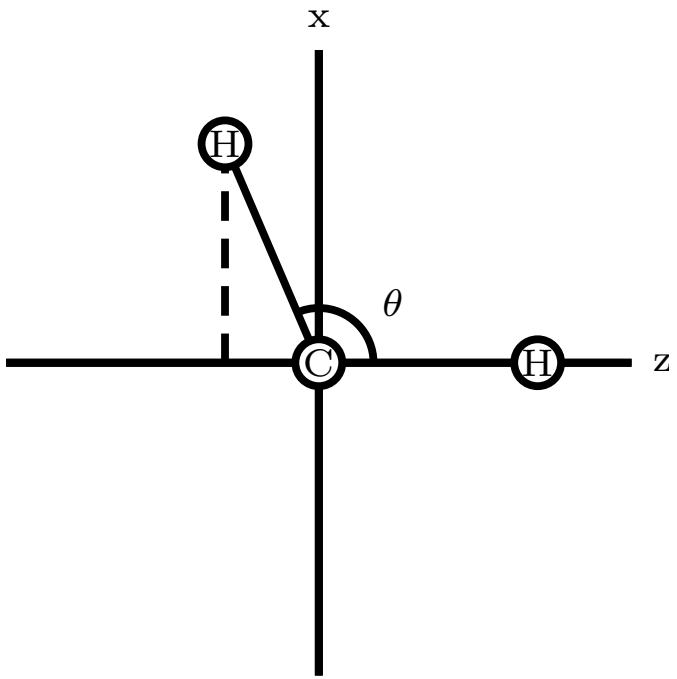
Methane Example

- The first atom, a, represents the origin of the coordinate systems. Set its three cartesian coordinates to $(0, 0, 0)$.
- The second atom, b, is at a fixed distance from the first one, which is their bond distance.
- Fix the z axis as the axis lying on the bond between the two atoms. b's coordinates are therefore $(0, 0, d)$ where d is the distance.
- The third atom, c, makes an angle with a and b and a bond with a.
- We can define the $x - z$ plane as the plane defined by the C atom and the first two hydrogens (every three non-collinear points define a plane).
- These two constraints and set the y coordinate to zero. Let r_{ac} be the distance between atoms a and c.

Methane Example

The x, z coordinates can be inferred by converting from polar to cartesian coordinates using the following formula:

$$z = r_{ac} * \cos(\theta) = -r_{ac} \cos(180 - \theta)$$
$$x = \sin(\theta) = r_{ac} \sin(180 - \theta)$$



Now that we have the x, z plane defined, the y axis can be extracted by a cross product, for example, between the two vectors $||b - a||$ and $||c - b||$, after normalization.

Methane Example

- If H3 were on the $x-z$ plane, it would make a 109.471° angle with C-H1 in the opposite direction, so its projection on the $x-z$ axis would be $\{1.027, 0.000, -0.363\}$.
- However, it has a dihedral angle of 120° with the C-H1-H2 plane, so we should rotate it by 60° around the z axis.
- z coordinate is unchanged, and x, y values are:

$$x = -1.027 * \cos(60) = -0.513$$
$$y = -1.027 * \sin(60) = -0.889$$

- Similarly, H4 creates a dihedral angle of -120° with the C-H1-H2 plane, so we rotate it by -60° around the z axis:

$$x = -1.027 * \cos(-60) = -0.513$$
$$y = -1.027 * \sin(-60) = 0.889$$

Methane Example

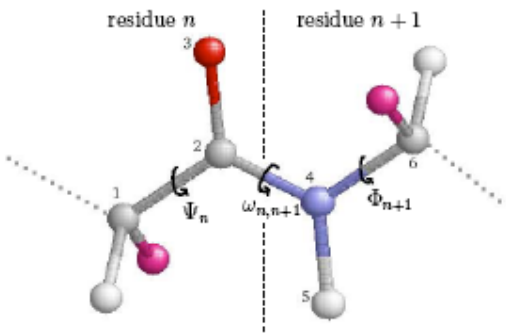
Re-orientating the molecule leads to Cartesian coordinates that make the symmetry more obvious:

Atom	X	Y	Z
C	0.000	0.000	0.000
H	0.629	0.629	0.629
H	-0.629	-0.629	0.629
H	-0.629	0.629	-0.629
H	0.629	-0.629	-0.629

Computational Modeling of Protein Folding

- What does the thermodynamic hypothesis suggest a naïve computer algorithm can do?
- The basic search involves enumeration of conformations
  - Exhaustive, brute-force search
- To be able to enumerate, the algorithm needs to have a way to first represent and then compute new conformations
  - Modeling, degrees of freedom
- To determine that some conformations are more relevant than others, the interatomic interactions need to be summed over to associate an energy estimate with each computer conformation
  - Energy function, scoring, ranking

Enumeration and Cyrus Levinthal's Paradox



- Consider a bond connecting two amino acids as the only source of flexibility
- Also consider that this bond can be in a limited number of configurations

Back-of-the-envelope calculation:

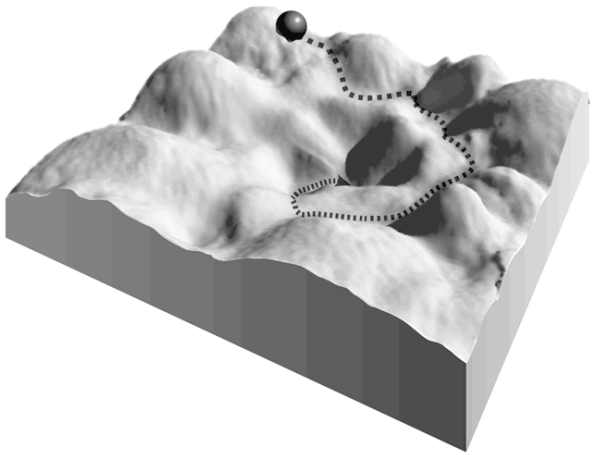
- A chain of 101 amino acids
- Has 100 such rotatable bonds
- Assume there are only 3 configurations available to each rotatable bond
- Then there are of 3^{100} conformations available to the protein chain

Enumeration and Cyrus Levinthal's Paradox

Levinthal's Paradox:

- Can a protein or algorithm enumerate 3^{100} conformations?
- Assuming a rate of 10^{13} conformations per second, it would take 10^{27} years for a protein sample all conformations
- Conclusion: This is NOT how proteins fold! [in the lab, proteins fold in milliseconds/microseconds]

The Non-paradoxical Levinthal's Paradox

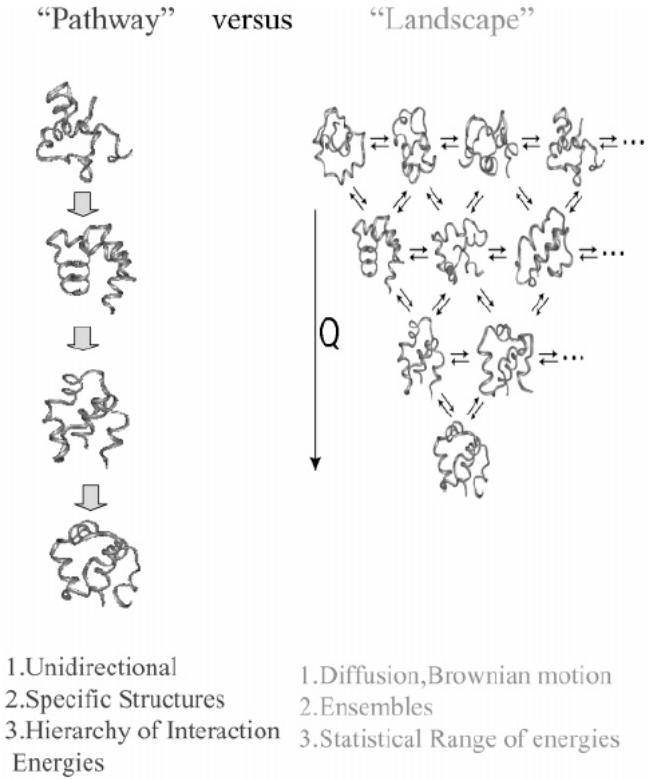


- Levinthal's paradox serves to illustrate that proteins do NOT sample conformations randomly and somehow stumble across the lowest-energy one
- There is method to the madness!

- Proteins fold faster than random trial suggests
- Then there must be an energetic bias that "guides" the protein towards the biologically-active conformation without going through unnecessary conformations
- What does the energy landscape look like?
- What does this mean for how proteins actually fold?

Pathway vs. Landscape

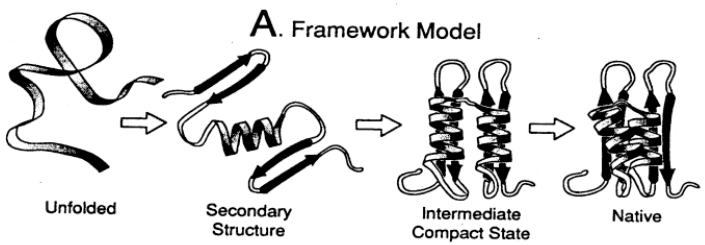
- Maybe there are specific pathways
- Information on which pathways the protein takes is encoded somehow in the unfolded states



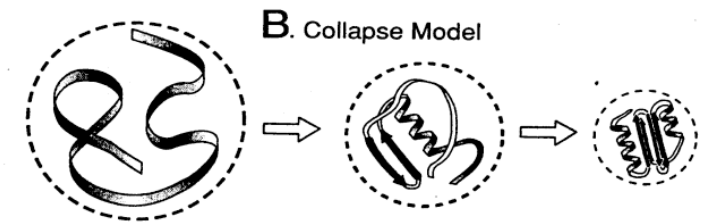
- Each state is seen as a conformational ensemble
- The energy landscape shows that the range of states becomes more limited with lower energy
- The landscape view encompasses the "preferred" pathways

Framework vs. Collapse

- Since proteins do not seem to fold at random, there must be an order of events that drives the folding of a protein chain.
- Important question: what drives the folding of a protein chain?



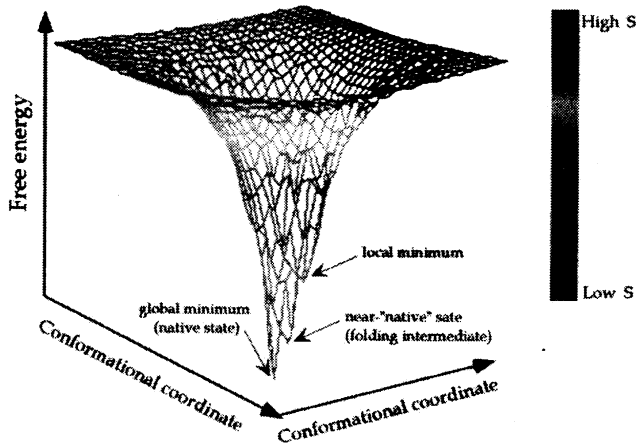
- Local interactions drive folding
- There is an order of events:
- Secondary structures form first
- Once formed, they stay that way
- The stable secondary structures then self-assemble in the native tertiary structure (folded conformation)



- Non-local interactions drive folding
- Proposed by Ken Dill in 1985
- Global collapse drives the secondary structure formation and not the reverse
- Folding code resides mainly in global patterns of contact interactions that are non-local

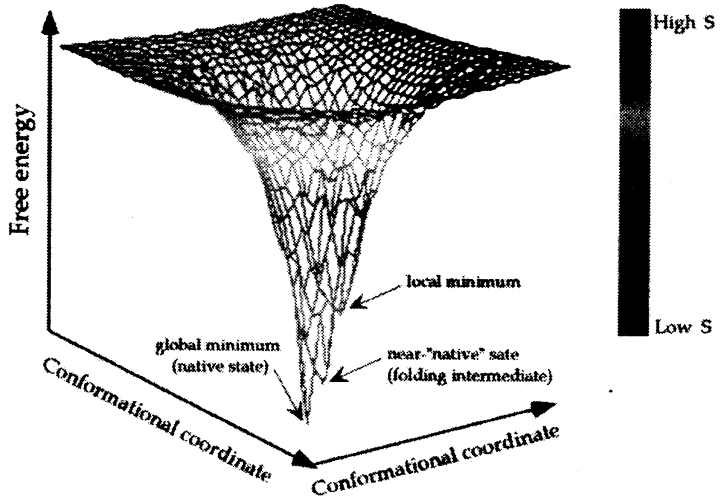
Funnel Landscape View of Protein Folding

- If proteins search for conformations, they do so on an energy landscape similar to a multidimensional funnel
- The slope of the funnel guides the protein down towards the energy minimum, where the native state resides
- The "folding funnel" leads to increase in rates of protein folding compared to expected rates for random diffusion processes
- The folding funnel also largely prevents entrapment of partially folded states
- The funnel view of protein folding was simultaneously proposed by Ken Dill and Peter Wolynes
- This energy landscape view of protein folding
  - Has been confirmed in experiment
  - Is now well accepted
  - Drives the logic/rules behind many computer algorithms that fold sequences in silico



The Landscape Encompasses the Pathways

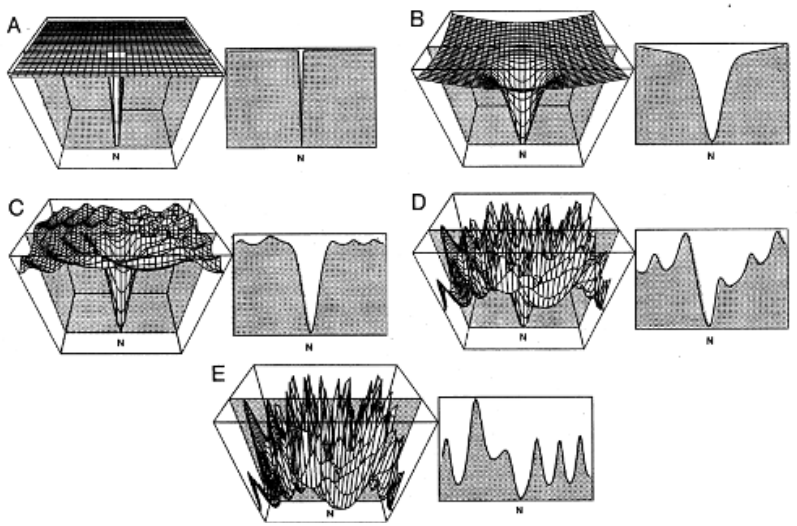
- As more sections of the native structure fall into place, the protein achieves lower energy, with the final folded native structure at the bottom of the funnel
- The landscape view allows thinking of multiple molecules rolling down the pathways of successive local minima until they all converge to the global minimum
- This process closely resembles the experiment, where measurements are obtained over multiple protein molecules – multiple replicas
- Width of the funnel (cross-section) gives the entropy – a measurement of how many different conformations achieve similar energy value
- As a protein folds, the free energy goes down – what is free energy?

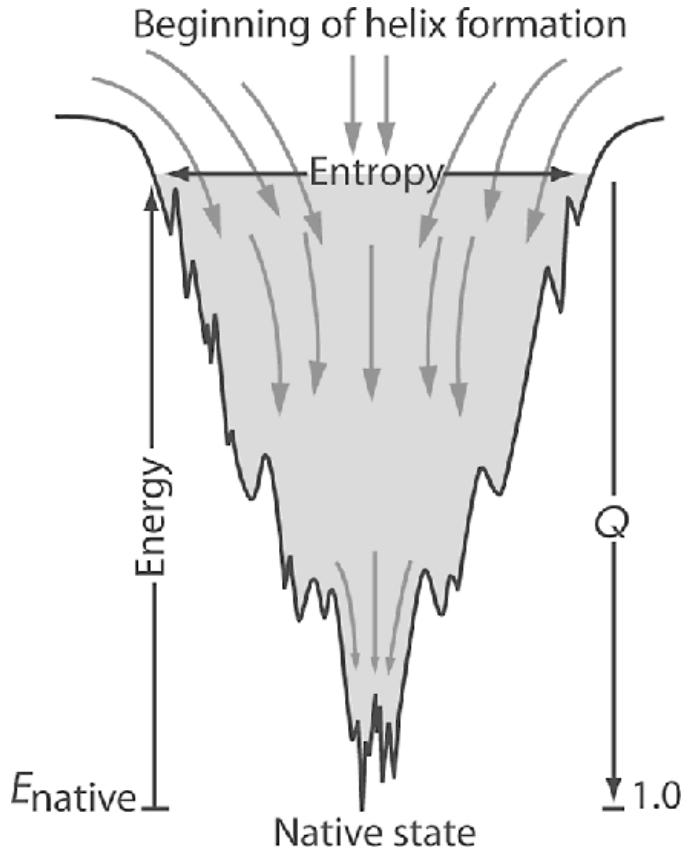


- The (Gibbs) free energy G = E - TS
- E = potential energy (interatomic interactions)
- T = temperature (on which folding happens)
- S = entropy (measures degeneracy of a state)

Landscape View and Assisted Folding

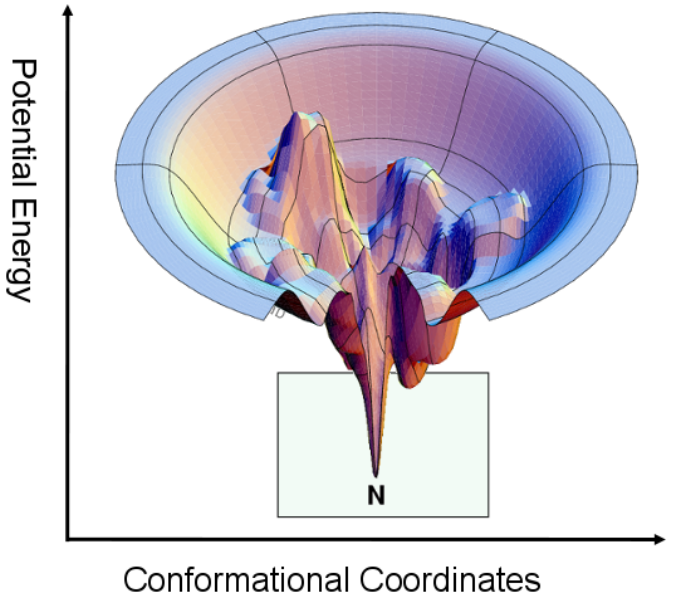
- Some experiments suggest that folding does not always lead to a unique state or structure corresponding to the overall free-energy minimum
- Kinetic hypothesis: a protein may get trapped in a local minimum in a really rugged energy landscape
- Some proteins are assisted in their folding by chaperones (other proteins that oversee and correct the folding process)





Folding With a Computer

- Protein folding tries to answer the how together with the what:
- Given a random conformation, how does the protein find its lowest-energy (native) state and what does this state look like
- Interested in both the final answer and in the actual manner the answer is obtained
- Protein Structure Determination addresses mainly the what:
- Given the amino-acid sequence, what is the native state (one structure, an entire ensemble of native-like conformations)?
- Computational methods that consider the how are categorized as folding methods.
- Methods that address the what are categorized as structure determination/prediction methods.



Challenges for a Computer Scientist

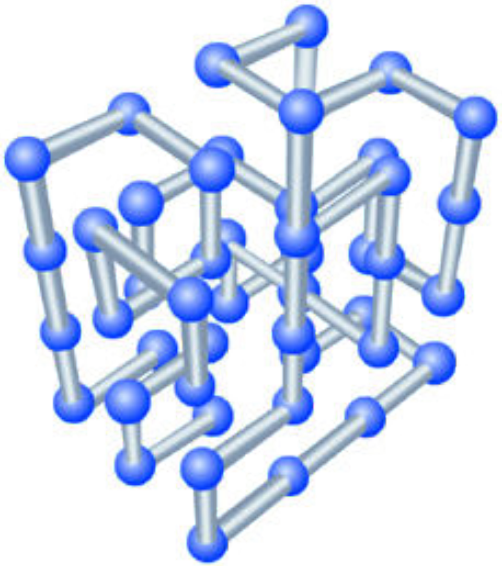
- Modeling – how are protein conformations represented
  - If all atoms are explicitly modeled and followed in space, this puts a lot of computational burden on a search algorithm. If not all atoms are explicitly modeled, how does one determine which ones to model?
  - Example: Focus on spatial arrangements of the backbone first
  - Are the modeled atoms allowed to move anywhere in space?
  - Example: Force atoms to be on a lattice – brings enumeration into the realms of feasible computation
  - Coarse-grained (not all atomic detail) vs. all-atom modeling is an important decision

Challenges for a Computer Scientist

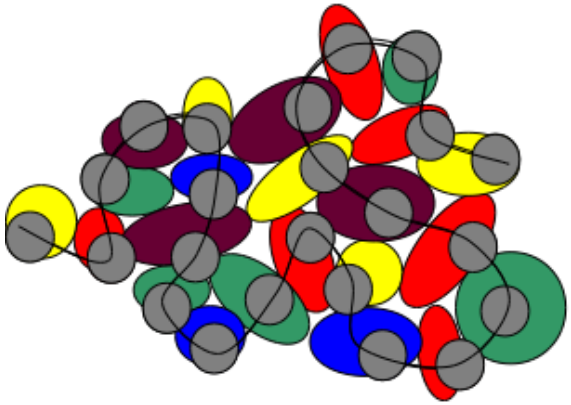
- Exploration/Search algorithm to traverse the conformational space
  - Trajectory-based exploration is one strategy to search the protein conformational space
  - Probabilistic Exploration that does not generate conformations in a trajectory is another
- Scoring – Energy function to determine whether a conformation is of low or high energy
  - Practical functions are empirical AND there are many of them
  - Very important to design functions that can accurately score coarse-grained conformations
  - Evaluation of an energy function on an all-atom representation is quite expensive

Representation of Protein Conformations

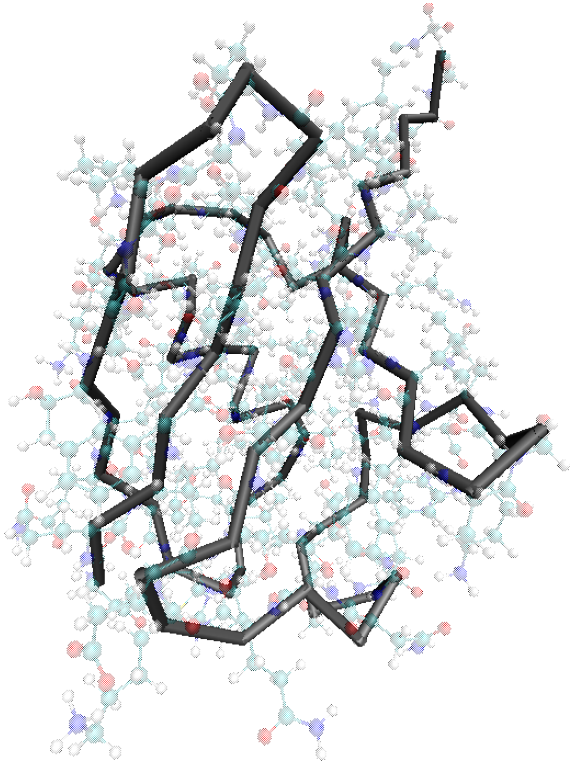
Discrete atoms in a lattice – realm of enumeration.



Continuous – off-lattice models.



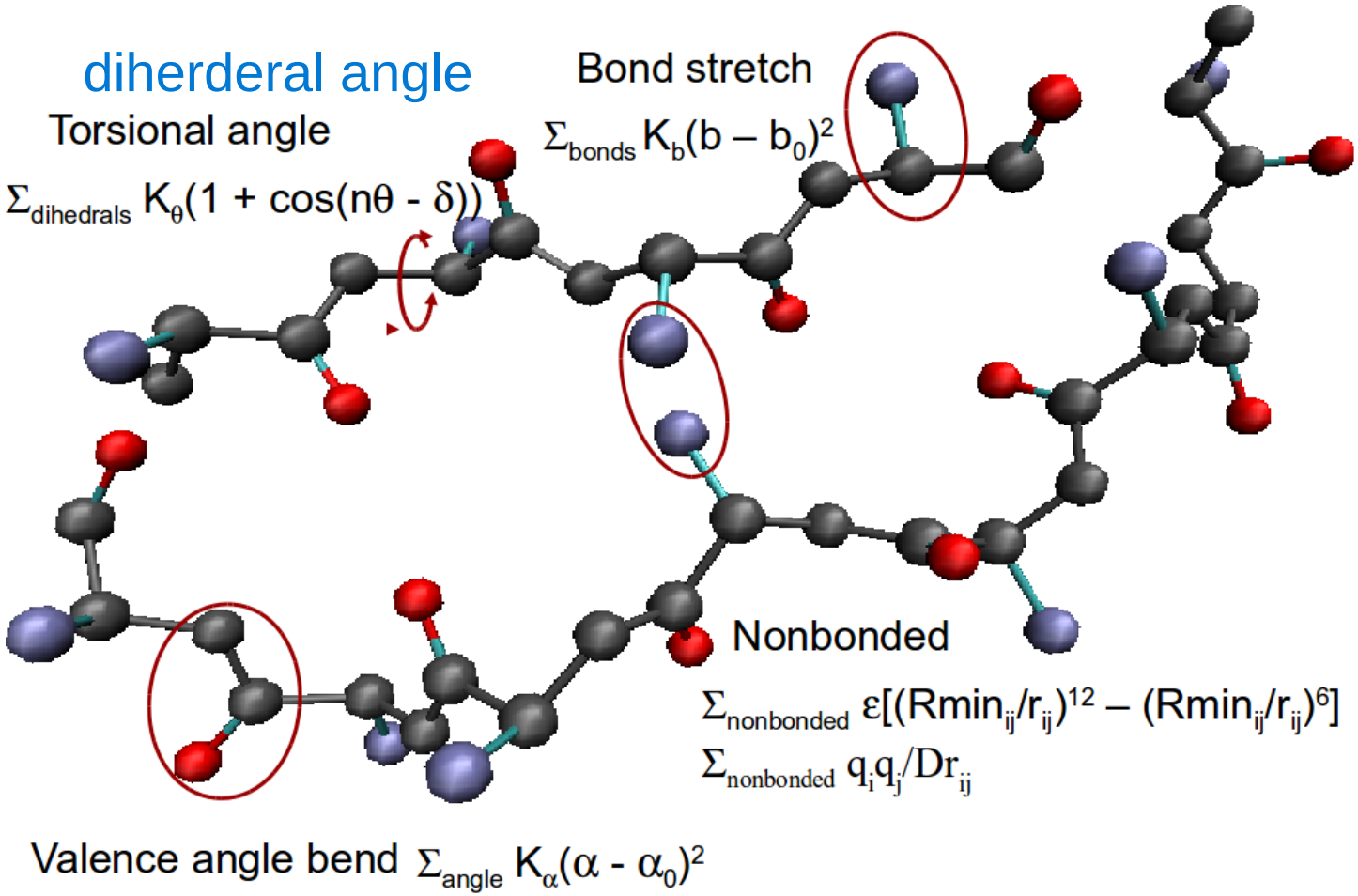
Coarse-grained modeling



Fine-grained modeling

Energy of Protein Conformations

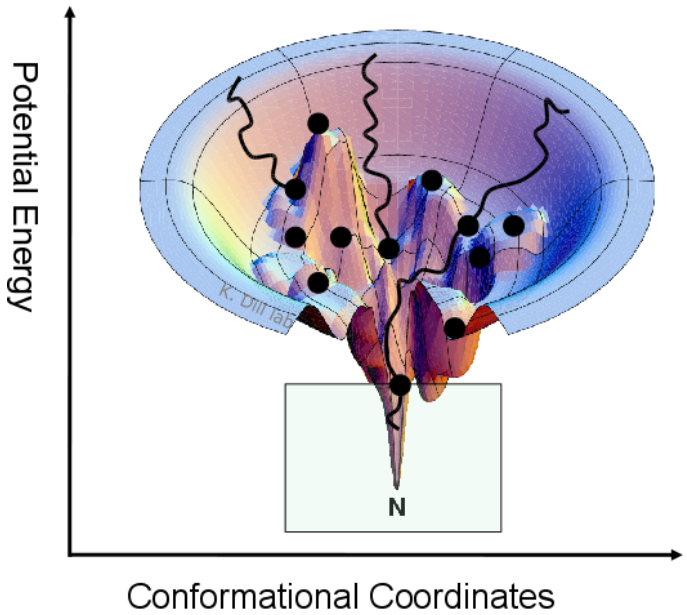
Empirical force-fields to measure potential energy AMBER ff*, CHARMM, OPLS, knowledge-based, ...



Launching Trajectories To Sample the Basin

- Trajectory-based exploration:
  1. Start with a conformation
  2. Generate a next one
  3. Continue until ...
- Systematic Search – Molecular Dynamics (MD)
- Probabilistic Walks – (Metropolis) Monte Carlo (MC)
- In MD, consecutive conformations in the trajectory are generated by following the gradient of the energy function.
- In MC, moves are sampled from a move set to generate the next conformation in the trajectory.
- The generated conformation is then accepted according to the Metropolis criterion.

f= ma mass we will get from above, a will be getting from atom



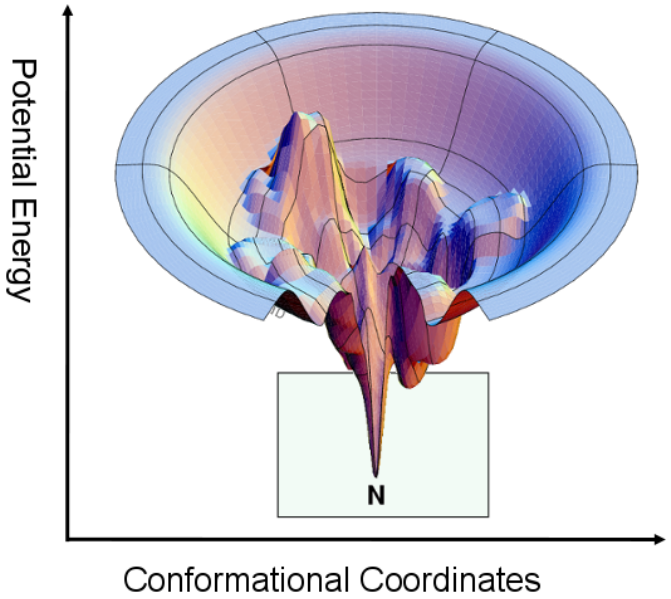
Trajectory-based Search of the Native State

Systematic Search: Molecular Dynamics (MD)

- Uses Newton's laws of motion
- Employs thermodynamics and statistical mechanics to simulate the folding process
- Detailed simulations follow the motion of each atom in space
- MD simulations demand a lot of computer time to simulate a few nanoseconds of the folding process
- Reveal detailed information about the folding process

The Conformational Space is Vast

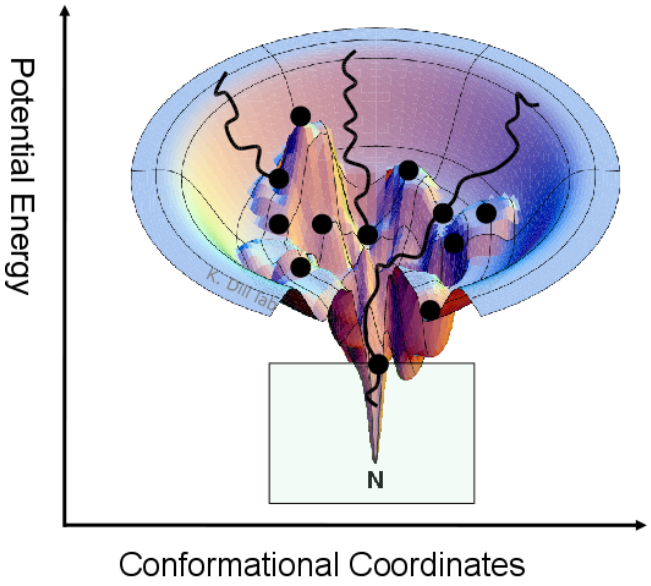
- Conformational space is high-dimensional: many atoms to follow in space
- Energy surface associated with conformational space [Onuchic J.N., Luthey-Schulten Z., and Wolynes P.G. Annu. Rev. Phys. Chem. 48, 1997]
- Evolution has "guided" native state (in naturally-occurring proteins) to be lowest free-energy state [Unger R. and Moult, J. Bull. Math. Biol. 55, 1993]



Locating the Basin of the Funnel with MD

1. evaluate forces on each atom
2. move atom according to force
   numerical update: $x(t + dt) = x(t) + dt * f(x(t)) + ...$
3. go back to 1.

- Numerical update is the bottleneck
- dt needs to be small for accuracy (1-2 fs)
- Generates a single trajectory
- Results depend on initial conditions, since MD is essentially a local optimization technique.



Trajectory-based Search of the Native State

Random (Probabilistic) Search: Monte Carlo (MC)

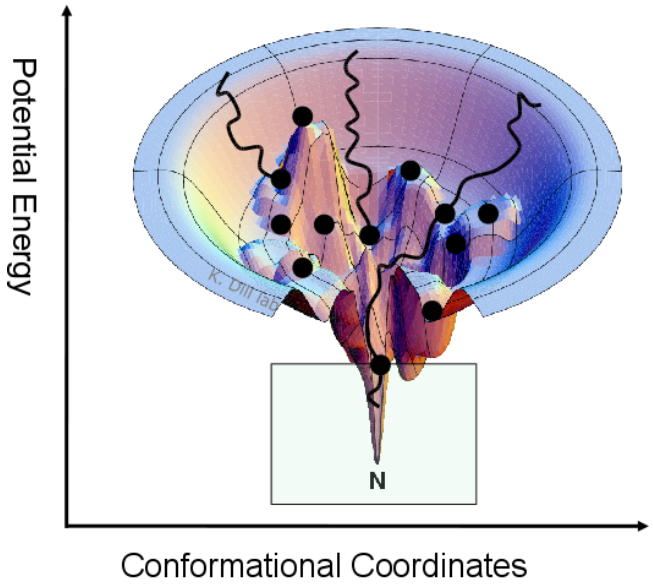
- Conducts biased probabilistic walks
- At each point in time (iteration), a move is made that is not necessarily physical or representative of what the protein does to transition between conformations
- A conformation resulting from a move is accepted with the Metropolis criterion based on its potential energy
- The paths taken to the native state may be actually impossible for a protein to follow
- By computing conformations through moves, MC exhibit higher sampling efficiency – they can make big jumps in conformational space

its like a drunken man walk every step will be on a new way

Locating the Basin of the Funnel with MC

Metropolis Monte-Carlo (MMC)

1. Start with a random/extended conformation $C_{curr} \leftarrow C_{start}$
2. Make a move (change C_{curr}) which results in a new conformation C_{new}
3. Give a value to one of the parameters considered example: rotate a bond
4. If $E(C_{new}) < E(C_{curr})$ then $C_{curr} \leftarrow C_{new}$
5. else $dE = E(Cnew) - E(Ccurr)$ $C_{curr} \leftarrow C_{new}$ with prob. $e^{-dE/scaling\_factor}$
6. Goto 2.

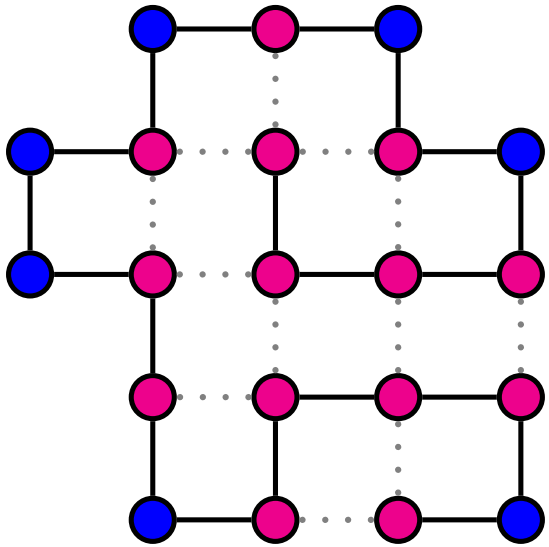


Pros and Cons of MD and MC

- Both MD and MC launch trajectories in conformational space
- The end-point of these trajectories depend to an extent on the initial conditions (conformations from which the trajectories were initiated)
- Both MD and MC are local optimization techniques aimed at sampling the global minimum in the energy landscape
- While MD gives physical trajectories, MC's trajectories may not correspond to a sequence of moves that a protein actually follows
- There may be nothing physical about the moves/parameters chosen to generate consecutive conformations
- This gives MC the ability to make bigger jumps in conformational space and sample the space faster than MD

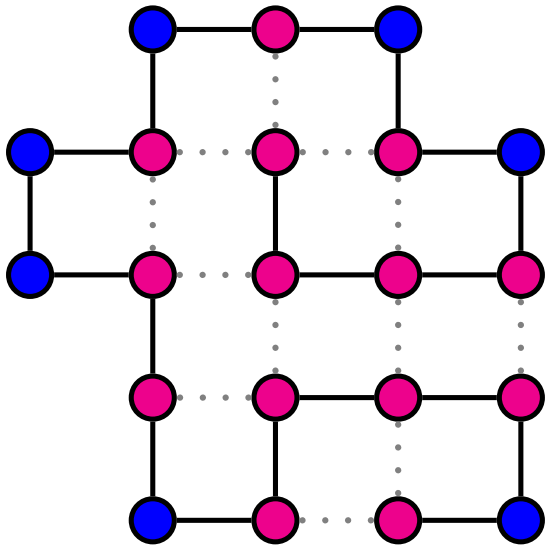
Lattice Models – The Simplest Model

- The conformational space is vast
- So, let's restrict positions in space for an amino acid to a lattice (cubic or diamond)
- This simple model allows to enumerate all possible conformations (on a short chain)
- The model can actually help answer the following question:
- **Possible question 1:** Which sequences lead to a stable native conformation? (protein design)



Lattice Models – The Simplest Model

- The conformational space is vast
- So, let's restrict positions in space for an amino acid to a lattice (cubic or diamond)
- This simple model allows to enumerate all possible conformations (on a short chain)
- The model can actually help answer the following question:
- **Possible question 2:** Given a sequence, what is its most stable configuration? (protein folding, structure prediction)



Lattice Based Models

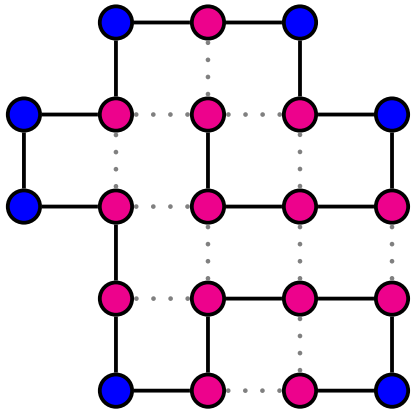
Potential course of action:

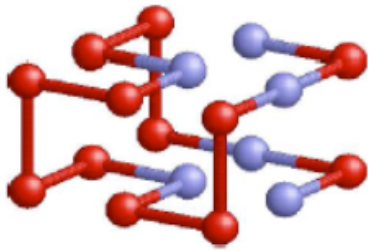
1. Enumerate possible amino-acid sequences
2. Twenty amino acids is too many, so we will have certain "representative" classes of amino acids
3. For each sequence, we will compute/enumerate the number of conformations on the lattice
4. We will then rank (through a ranking/scoring energy function) the enumerated conformations to determine the good folds

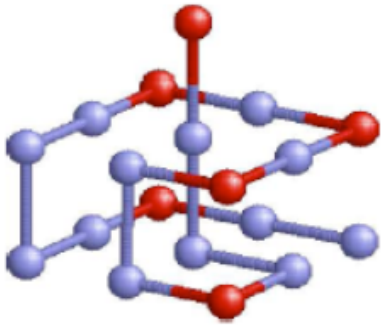
Hopefully we will find out which sequences yield good folds

HP Lattice Model: Ranking Sequences and Folds

- There are only two types of amino acids – H (red) for hydrophobic, P (blue) for polar (hydrophilic)
- H units repel water, P units attract water
- The energetic forces acting between the units are reduced to a single rule:
  - H's like to stick together (P units are inert, neither attracting nor repelling) This is a very simplified model.
  - The chain folds on a lattice laid out on graph paper.
- H's and P's are placed at the grid points of the 2D lattice
- The peptide bonds are lines drawn on the grid
- Confinement on the lattice keeps the number of conformations finite (amenable to enumeration)
- Several lattice geometries are possible in both 2D and 3D

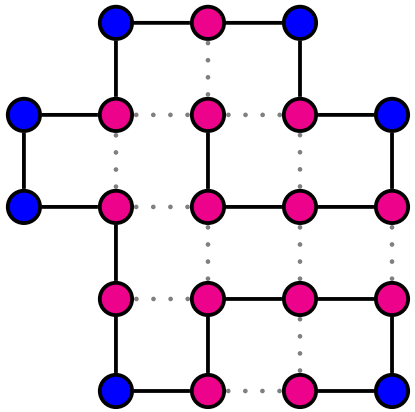






HP Lattice Model: Self-avoiding Walks

- We can ask ourselves two related questions:
- Given an n-aa chain of H's and P's – How many sequences are there?
- Given a specific sequence of H's and P's – How many conformations are there?
- This translates to the number of self-avoiding walks on the 2D lattice on the right
- A self-avoiding walk is a path on the 2D grid that does not cross the same grid point more than once

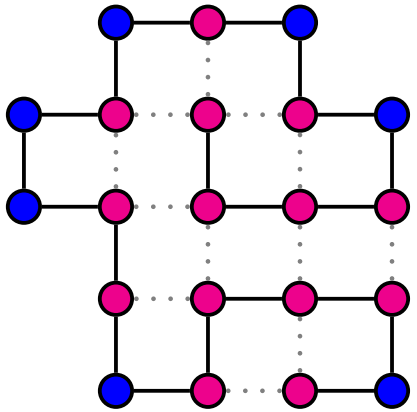


Remember course of action:

- Enumerate all possible HP sequences
- For each sequence, enumerate the number of conformations (self-avoiding walks) on the 2D lattice
- Collect statistics on which sequences give good folds

HP Lattice Model: Self-avoiding Walks

- How many shortest self-avoiding walks of length n-1 are there on the 2D grid?
- Walks of length 1: four possible directions: North, West, South, East [4 walks]
- Walks of length 2: From the ends of the walk of length 1, three choices (remember self-avoiding) [3 x 4 = 12 walks]
- Walks of length 3: 36
- ... Walks of length 51: 10^{22}



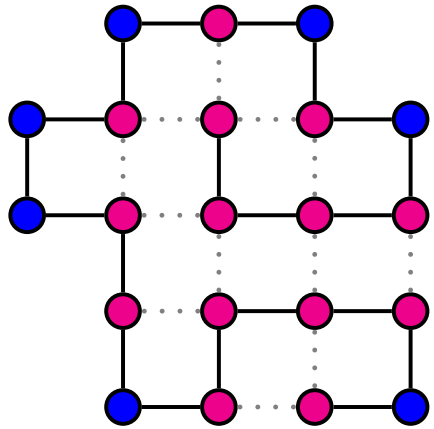
Remember course of action:

- Enumerate all possible HP sequences
- For each sequence, enumerate the number of conformations (self-avoiding walks) on the 2D lattice
- Collect statistics on which sequences give good folds

HP Lattice Model: Good and Bad Folds

What makes a fold good?

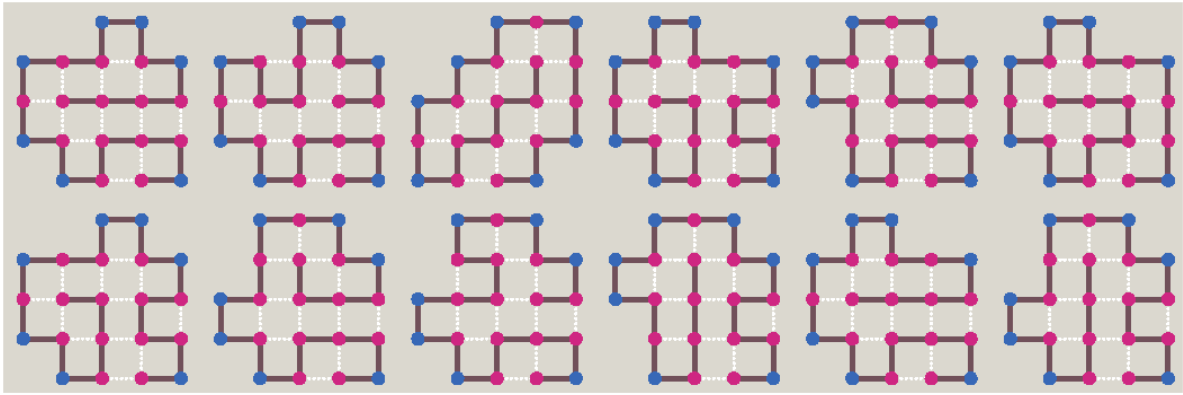
- Low Gibbs free energy G = E - TS
- Low potential energy, high entropy
- An extended chain has high E but low S
- E is high because amino acids that "want" to be close together are forced to be apart
- S is low because the extended chain is highly ordered
- Folding the chain should bring it back into a shape with a low overall value for G



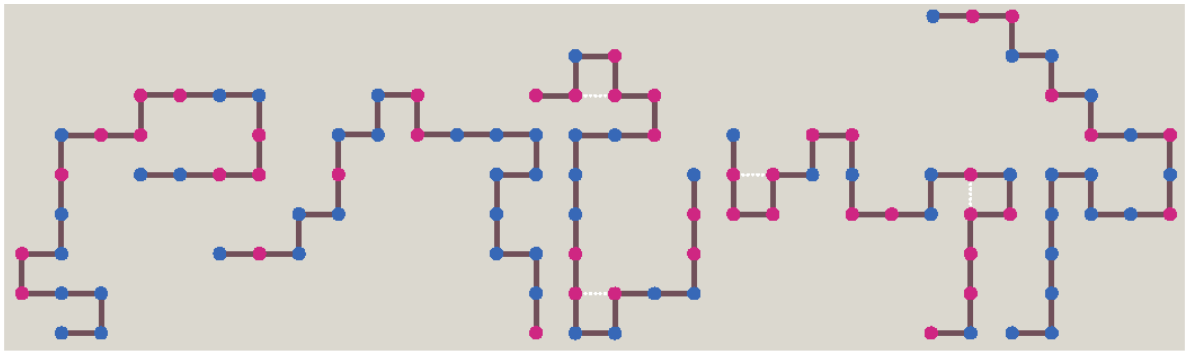
How do we model G here?

- Approximation: count only the number of H-H contacts (dashed lines)
- This should reflect the tendency of hydrophobic amino acids to go towards the core of the protein to avoid water
- Low G is now maximum number of H-H contacts – this is our scoring function

Finding Best 2D Lattice Folds of HP sequences



12 of the 107 most stable folds of 80 21-aa sequences: Red H's form stabilizing contacts (dotted white lines) when nearest neighbors blue P beads have no interactions (not counted in the G function)



There are 117,676,504,514,560 possible folds of 21-aa sequences. The 5 shown here are selected at random.

Monte Carlo to Sample the Conformational Space

- On long chains, enumeration is impractical: Monte Carlo could compute self-avoiding walks in the lattice
- Given an HP chain of length N
- Start with a random self-avoiding walk on the $N \times N$ grid
- For i $= 1$ to N_{cycles}
  - Propose a move to obtain a new self-avoiding walk from the current one
  - Estimate the G value of the proposed walk with the ranking H-H score and compare it to the G value of the current walk
  - If the difference in energy meets the Metropolis criterion accept the new walk and resume the next cycle from it

Monte Carlo on Off-lattice models

- Lattice models are a crude approximation of the real conformational space of proteins
- How would you apply Monte Carlo to a backbone chain of atoms in 3D (off-lattice models)?
- Your parameters (movable set) are the ϕ, ψ angles of the chain
- Where would you draw values from to attempt moves?
  - Suggestion 1: random angle values from -π to π
  - Suggestion 2: angle values sampled from the Ramachandran map of an amino-acid
- What energy function would you implement to score the new conformations?
  - Suggestion 1: minimize the number of collisions - crude approximation. Often works, not very accurate.
  - Suggestion 2: also reward atoms in specific distances from each-other (van der Waals)

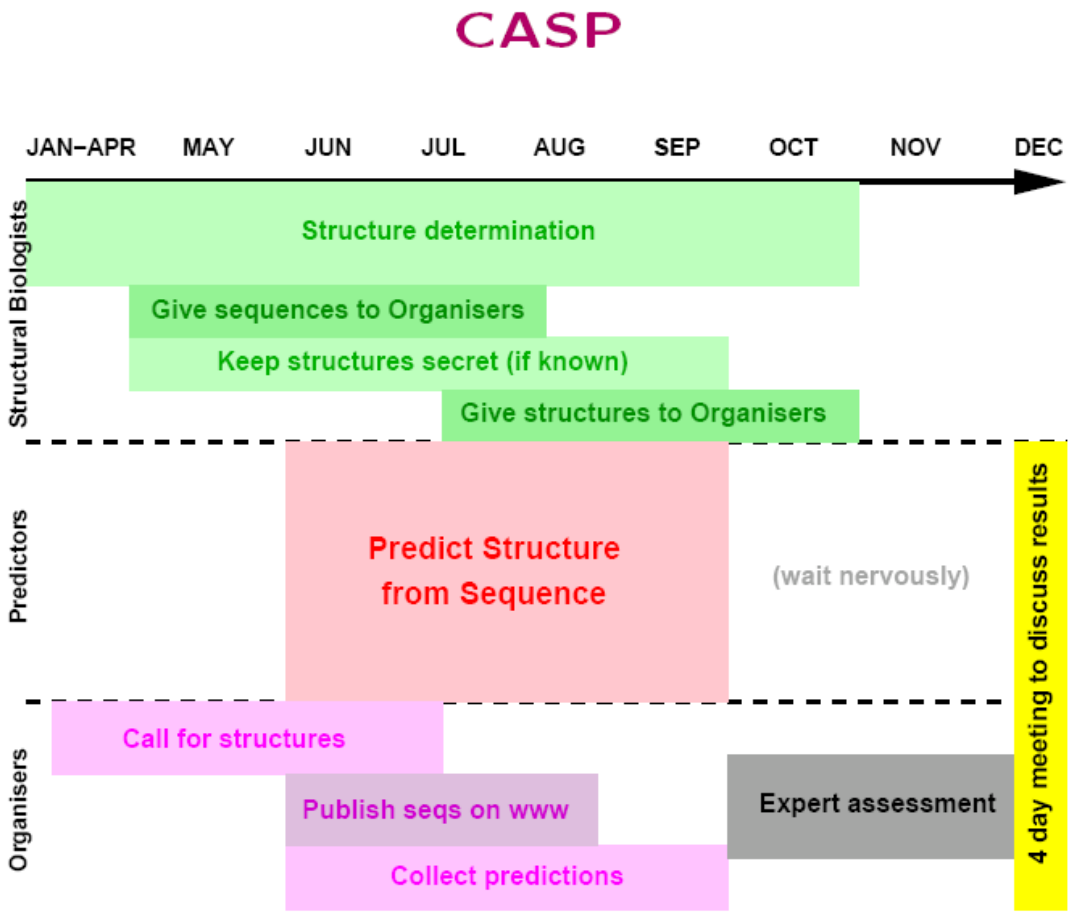
Methods for Structure Prediction and CASP

- Problems
  - Secondary structure prediction
  - Fold recognition
  - Tertiary structure prediction
- Methods
  - Knowledge-based methods for secondary structure prediction
  - Homology modeling for fold recognition and tertiary structure prediction
  - Ab-initio or physics-based methods for tertiary structure prediction
  - Hybrid methods combine knowledge and physics for tertiary structure prediction
- Quality assessment
  - For tertiary structures: RMSD, IRMSD, LGA
  - For binding sites: Shapes, Surfaces, Cavities

CASP: History and Motivation

- Critical Assessment of techniques for protein Structure Prediction (CASP)
- First held in 1994 on a biennial basis – 9 meetings up to 2010
- Protein Structure Prediction Center: `http://predictioncenter.org/`
- A competition that pitches computational groups against one another on their solutions to the structure prediction problem on various proteins (whose experimental structures have just become available but are temporarily withheld from deposition in the PDB)
- Problems addressed since 1994 (some of them discontinued due to great success):
  - Secondary Structure Prediction – discontinued due to high accuracy
  - Fold Prediction/Remote Homology Detection Method
  - Tertiary Structure Prediction

CASP – The Process



CASP – List of Addressed Problems

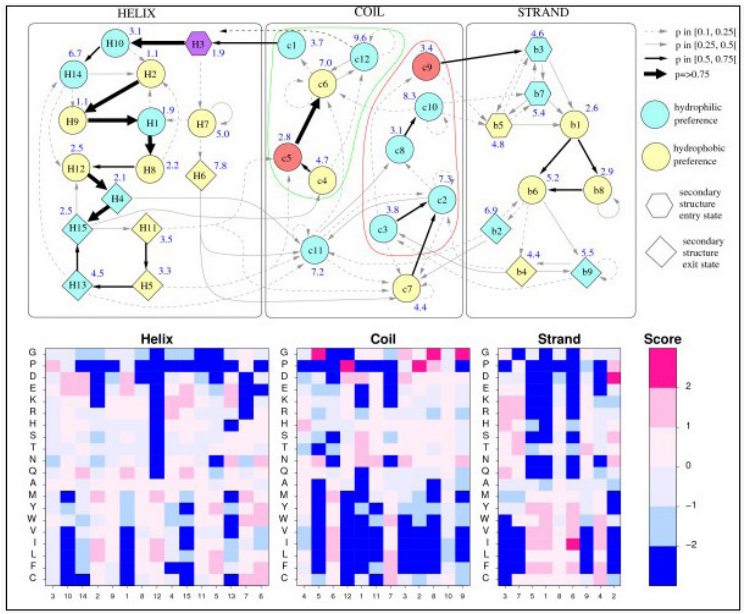
- Domain prediction: Identify regions that fold into "independent" compact structures
- Secondary structure prediction: Identify local structural regions of the protein
- Contact-map prediction: Identify residue pairs of residues in "native contacts"
- Fold recognition: Determine whether the protein adopts a shape that is already known
- New-fold prediction: Determine the tertiary structure of a protein with an unknown shape
- Function Prediction: Determine the functional class of the protein
- Model Quality Assessment: How well each group does

CASP: Pushes State of In-silico Research

- Human experts are allowed to have their submissions assessed and ranked
- The focus of CASP, however, is on assessing computational approaches and methods
- Another problem considered in CASP competitions is fold recognition or prediction
  - Reason: even if one cannot obtain the precise details of the tertiary 3D structure, one can at least recognize the fold from the sequence
  - Use: tertiary structure details are easier to compute once the overall fold is known
- The main interesting problem in the last years is prediction of tertiary structure from the amino-acid sequence

Secondary Structure Prediction: State of the Art

- Generative Models: learn a model based on specific signals and see how well the model explains (classifies/annotates) candidate queries
- Bayesian Methods, HMM
- Example: OSS-HMM with accuracy 75.5%
- Discriminative Models: Key idea is to separate between the various classes or groups
- SVM, Neural Net, Logistic Regression
- Example: YASSPP uses SVM-based models for each C(oil), E(xtended), H(elix) state



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1769381/

Secondary Structure Prediction: State of the Art

Summary of neural networks for proteins secondary structure prediction.

Method	Ref.	Year	Acc (%)	Dataset	Description
DL	[50]	2015	Q3 = 80.7%, SOV = 74.2%	CASP (CASP9: 105 proteins, CASP10: 93 proteins)	deep learning (belief) network; PSSM by PSI-BLAST
	[51]	2015	Q3 = 81.8%	CASP11	local backbone angles; PSSM;physical chemical properties; deep learning neural network
	[52]	2016	Q3 = 84.7%, SOV = 86.5%, Q8 = 72.3%;	CASP;	deep convolutional neural fields; conditional neural fields (CNF); PSSM; http://raptorx.uchicago.edu/
			Q3 = 82.3%, SOV = 84.8%, Q8 = 68.3%	CB513	download/;
	[53]	2016	Q3 = 82.9%, Q8 = 68.20%	CB513;	PSSM, recurrent neural networks, encoder–decoder networks, bidirectional
			Q3 = 84.2%,Q8 = 73.1%	CullPDB	gated recurrent units
RNN	[59]	2006	Q3 = 73.1%, SOV = 63.0%	CB396 (RS126 set for training)	bidirectional segmented-memory recurrent neural network; dynamics; multiple alignment profile generated from BLAST
	[55]	2007	Q3 = 74.38%, SOV = 66.05%	PSIPRED (training dataset:EVA)	cascaded bidirectional recurrent neural networks, long-range interactions; strong correlation; PSSM
	[58]	2010	Q3 = 79.36%, SOV = 70.09%.	PSIPRED	bidirectional recurrent neural network; reciprocal recurrent neural network, long-range interactions; strong correlations; PSSM;
	[56]	2013	Q3 = 82.2%	1630 proteins of lower quality form PDB	relative solvent accessibility; cascaded architecture; http://distillf.ucd.ie/porterpaleale/
	[60]	2017	Q3 = 83.9%	TS115	BRNN, PSSM, DSSP
RBF	[25]	2005	Q3 = 77.4%	340 protein from PDB	conformational classification; structure transition
	[63]	2008	Q3 = 71.3%	RS126	two level RBFNN; evolutionary information
other neural networks	[71]	2007	Q3 = 72.01%	CB513 data	PSSM by PSI-BLAST; parallelization; Pthread and OpenMP; BLOSUM62; tertiary classifier
	[64]	2008	Q3 = 73.4%	RS126	two-stage architecture; fully connected multilayer perceptrons (MLP) neural network; backpropagation algorithm; Sequence Profiles; http://raptorx.uchicago.edu/download/
	[72]	2011	Q3 = 81.8%; Q3 = 82.0%	CASP9; SPINE (Sub-dataset of PISCES)	multistep NN model; torsion angle prediction; solvent accessible surface area;
	[67]	2013	Q3 = 82.14%	CB513	three layered of CVNN; circular transformation; energy potentials; CABS algorithm;
	[68]	2016	Q3 = 81.72%	CB513	heuristics; Complex-valued relaxation network; inhibitor peptides; compact model; energies computed

https://pubmed.ncbi.nlm.nih.gov/28763690/

CASP: Fold Recognition

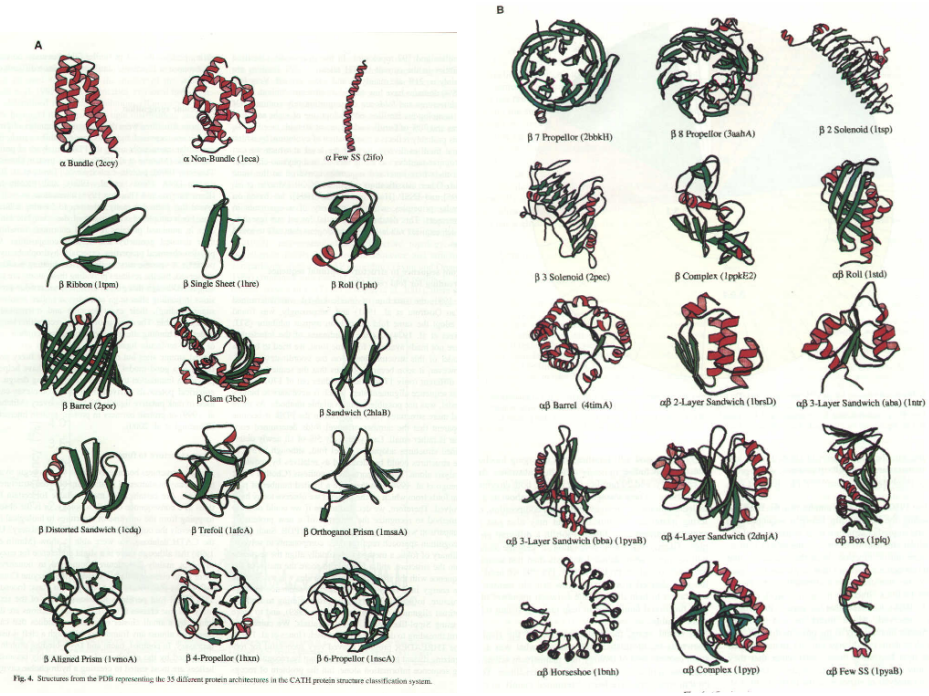
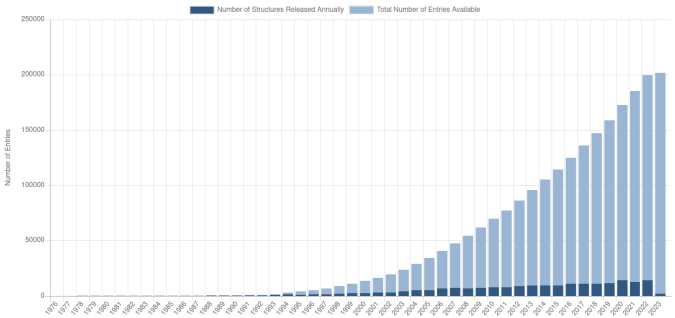
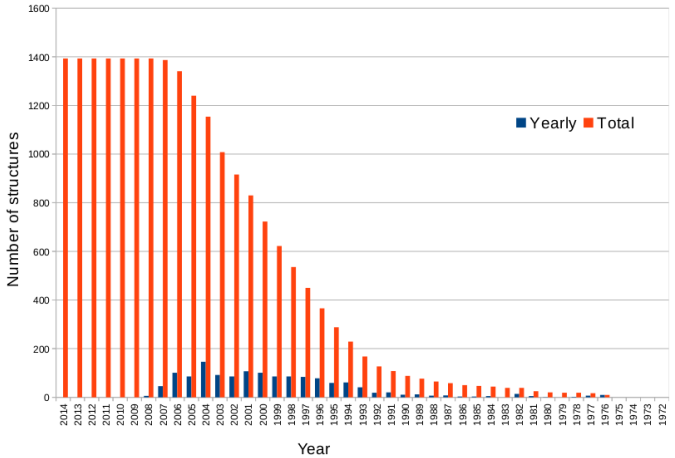


Fig. 4. Structures from the PDB representing the 35 different protein architectures in the CATH protein structure classification system.



http://www.rcsb.org (PDB)



New "fold" discovery is decreasing

Approaches to In-silico Structure Prediction

- **Homology modeling:** use information from known structure(s) with similar sequence(s)
  - Main idea: knowledge-based approach, exploiting databases of known protein structures
  - Applicability: if $> 60\%$ sequence identity, problem is (many times...) solved
  - Challenge: $< 30\%$ sequence identity, poor results
  - Challenge: Loop regions are usually not very well conserved
  - Good for sequences that have close homologs with known structures.

Approaches to In-silico Structure Prediction

- **Ab-initio structure prediction:** predict without prior knowledge of other structures
  - Main idea: conformational sampling/searching combined with optimization of an energy function
  - Conformational sampling includes trajectory-based exploration, enhanced sampling, robotics-inspired methods, and more
  - Energy function can be physics-based (CHARMM, AMBER), knowledge-based (obtained from statistics over database of known structures), or have terms of both
  - Applicability: on proteins with $< 30\%$ sequence identity (or more...)
  - Challenge: trajectory-based exploration methods may take too long

Approaches to In-silico Structure Prediction

- **Hybrid methods:** combine ingredients from knowledge-based and ab-initio methods
  - Hierarchical approaches, fragment-based assembly dominate at CASP
  - Applicability: where just knowledge-based or just ab-initio are inadequate
  - Challenge: extend applicability to longer protein chains and multi-chain complexes
  - Even most sophisticated methods are still not very accurate.
  - Recently, deep learning has made significant progress (AlphaFold + other learning based methods for contact prediction)