CS2220: Introduction to Computational Biology

Lecture 7: Protein Structure and Optimization

Lisa Tucker-Kellogg

11 March 2010

NUS National University of Singapore

Contents

- **Protein Structure**
- **Molecular Modeling and Potential Energy**
  - Mainly in the context of Proteins
- **Optimization Methods**
  - Simulated annealing
- **Optimizing Pseudo-Energy**
  - NMR structure determination

Protein Structure

Primary Structure

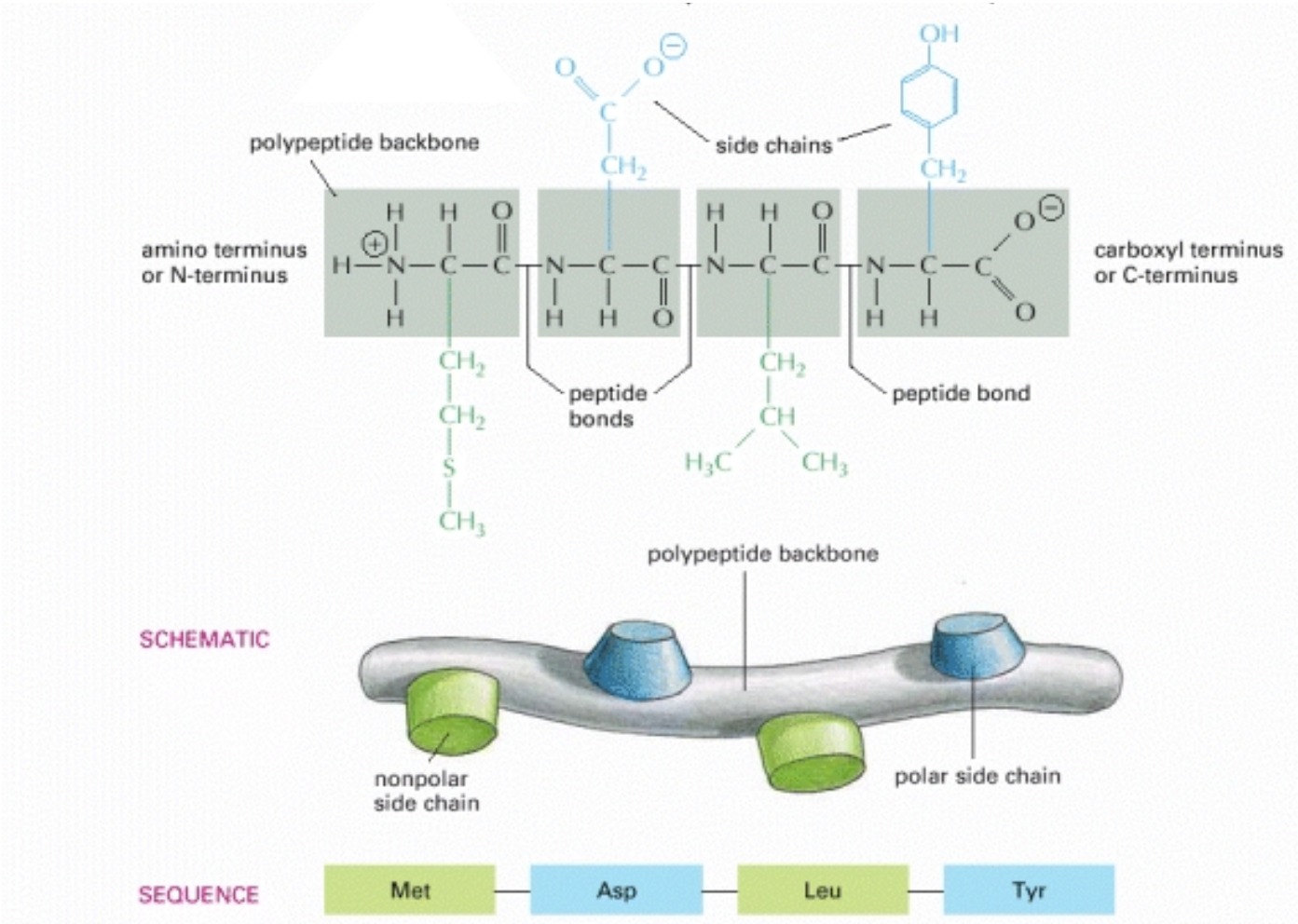


Image credit: Alberts et al., Mol Biol Cell

Rotational Freedom of Backbone

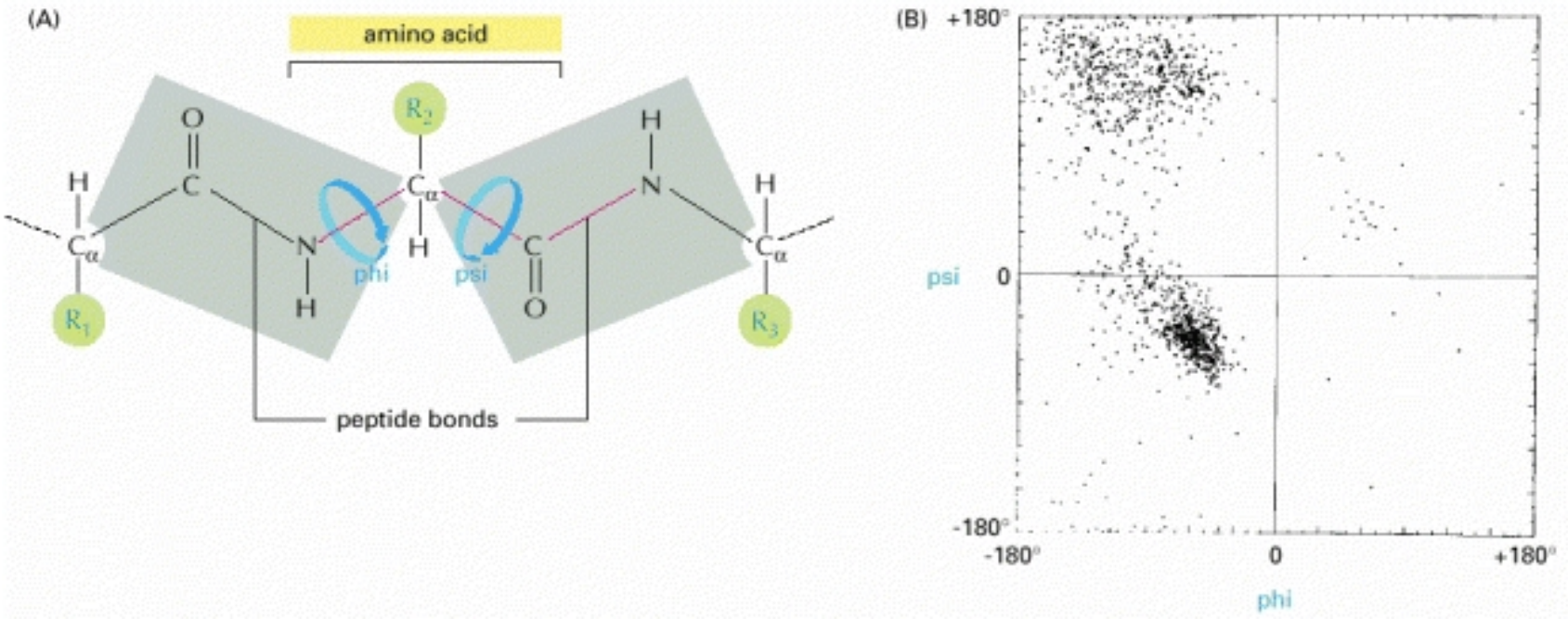


Image credit: Alberts et al., Mol Biol Cell

Backbone Makes Hydrogen Bonds

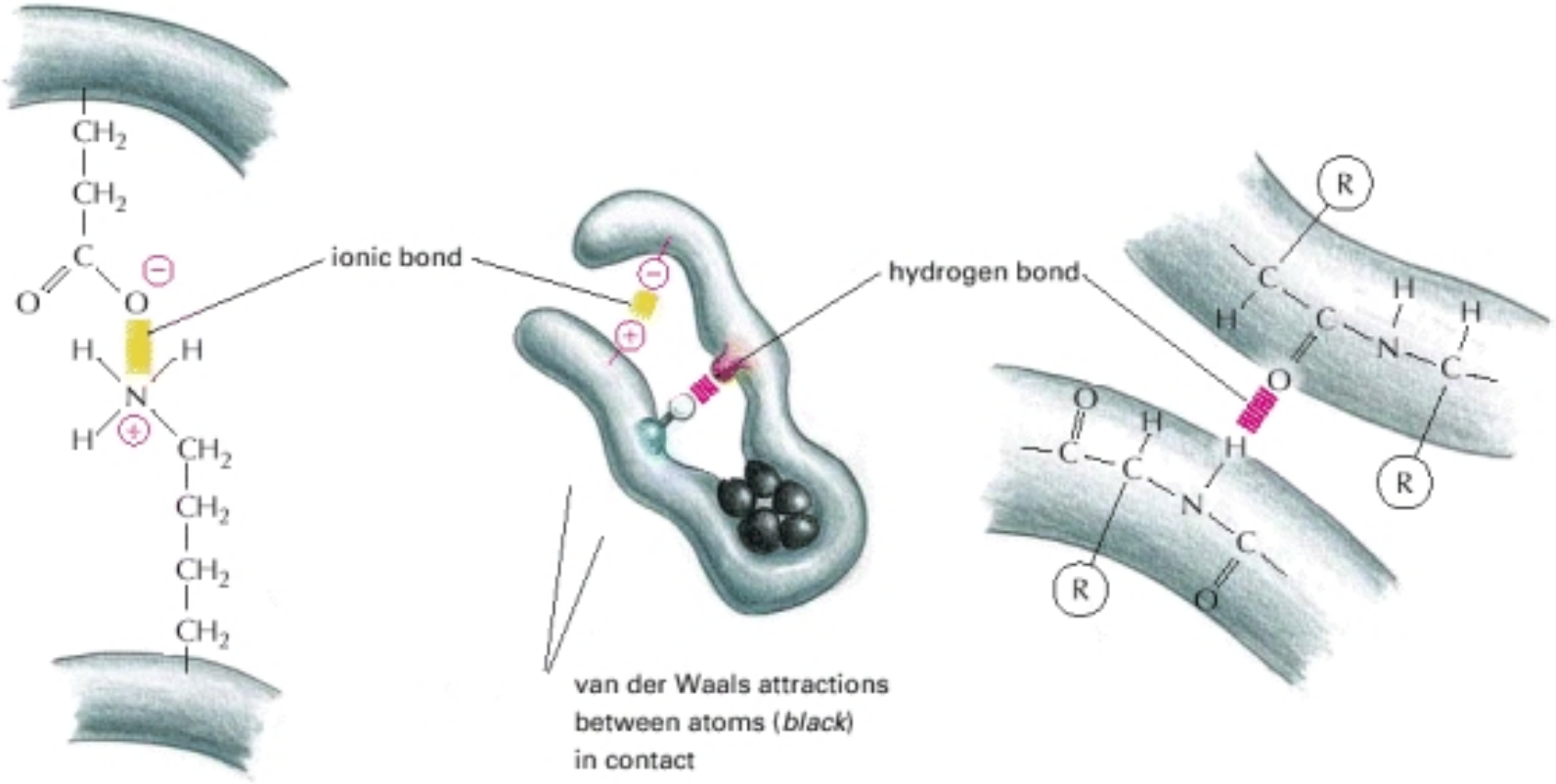


Image credit: Alberts et al., Mol Biol Cell

A hydrogen bond is NOT a covalent bond with a hydrogen

Protein Secondary Structure alpha-helices and beta-strands

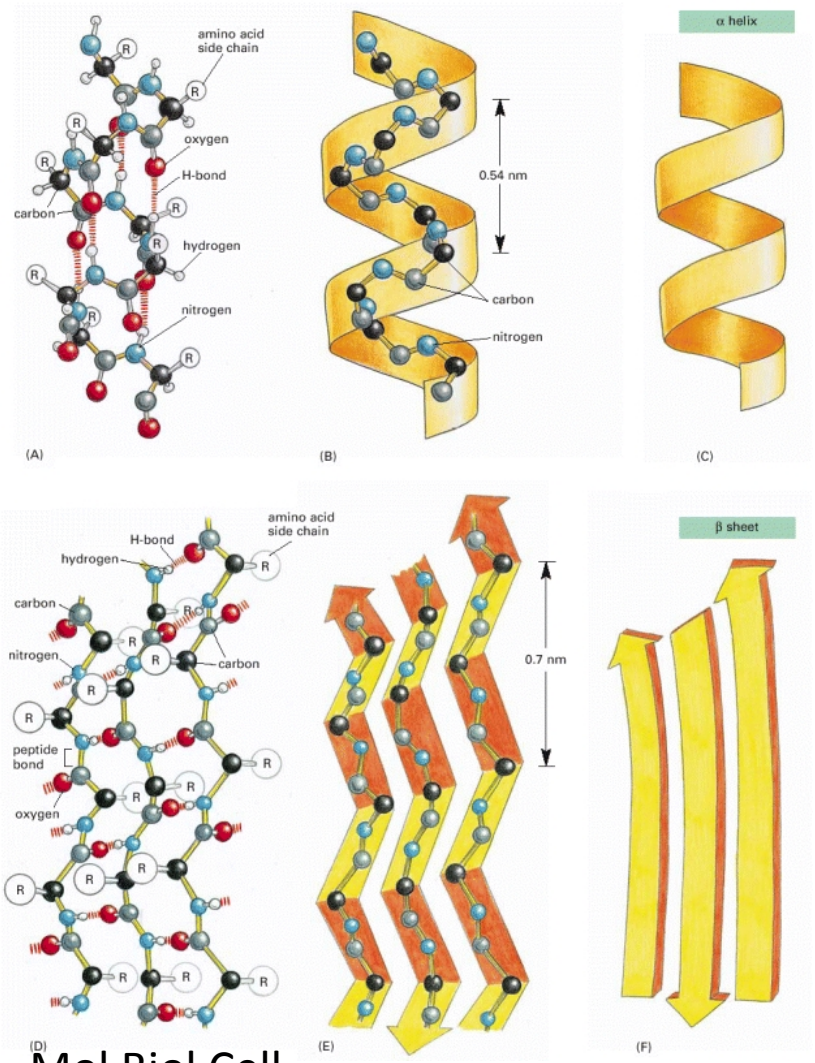


Image credit: Alberts et al., Mol Biol Cell

Anti-Parallel and Parallel Beta-Sheets

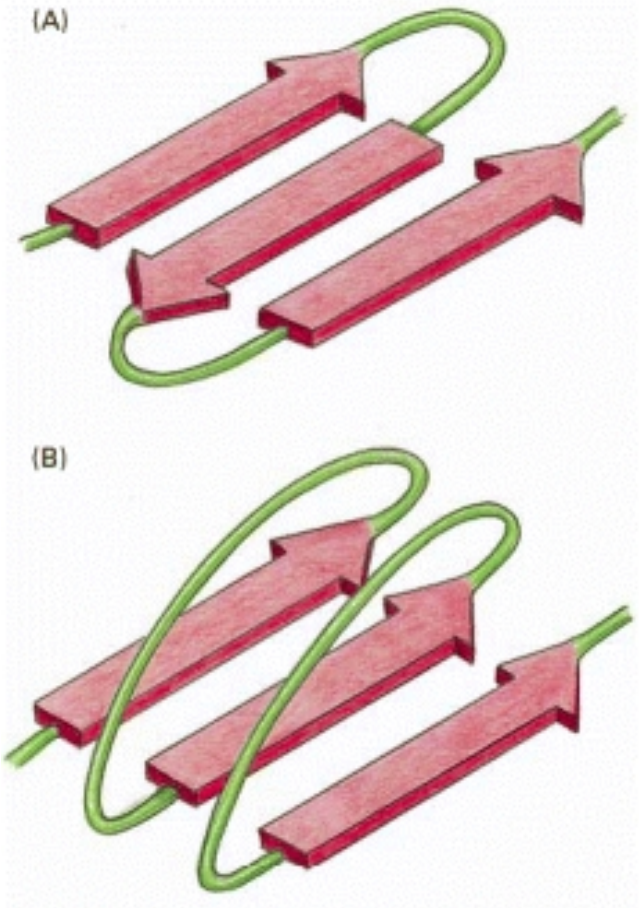


Image credit: Alberts et al., Mol Biol Cell

RNA has secondary structure too

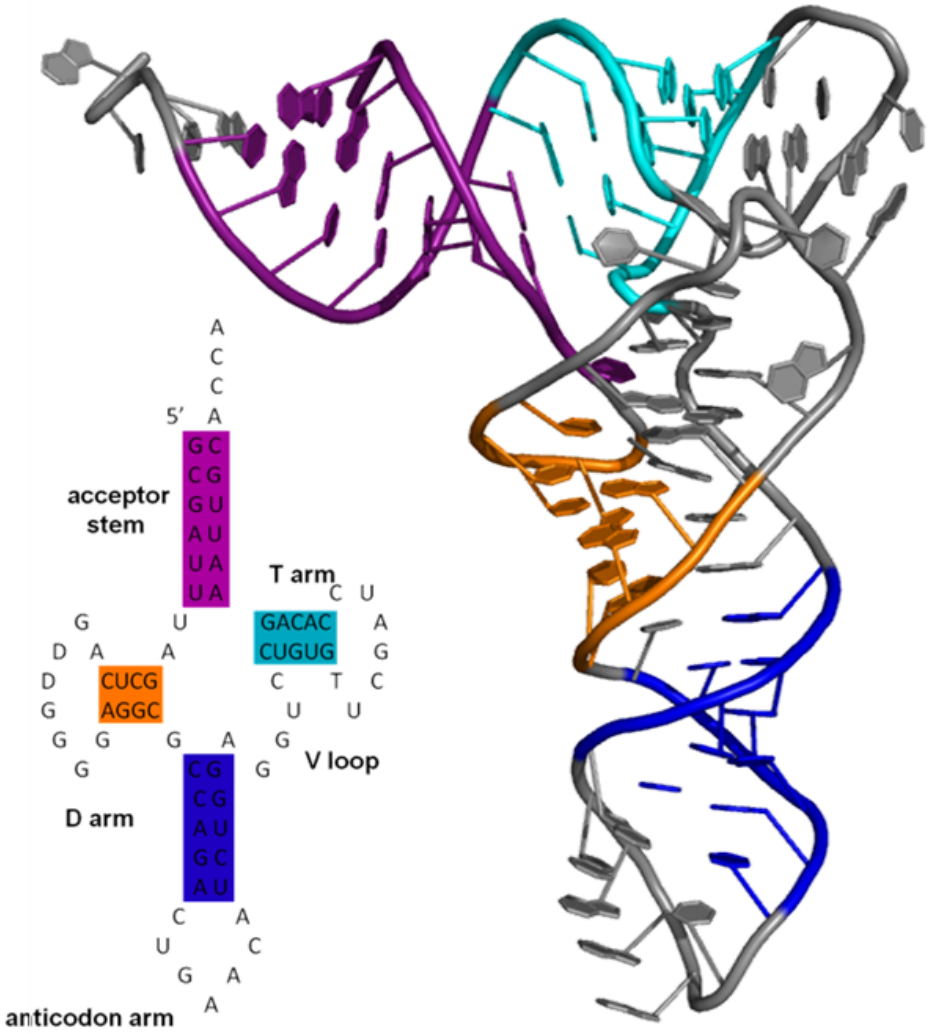


Image credit: Wikipedia

Hydrophobic and Hydrophilic Groups

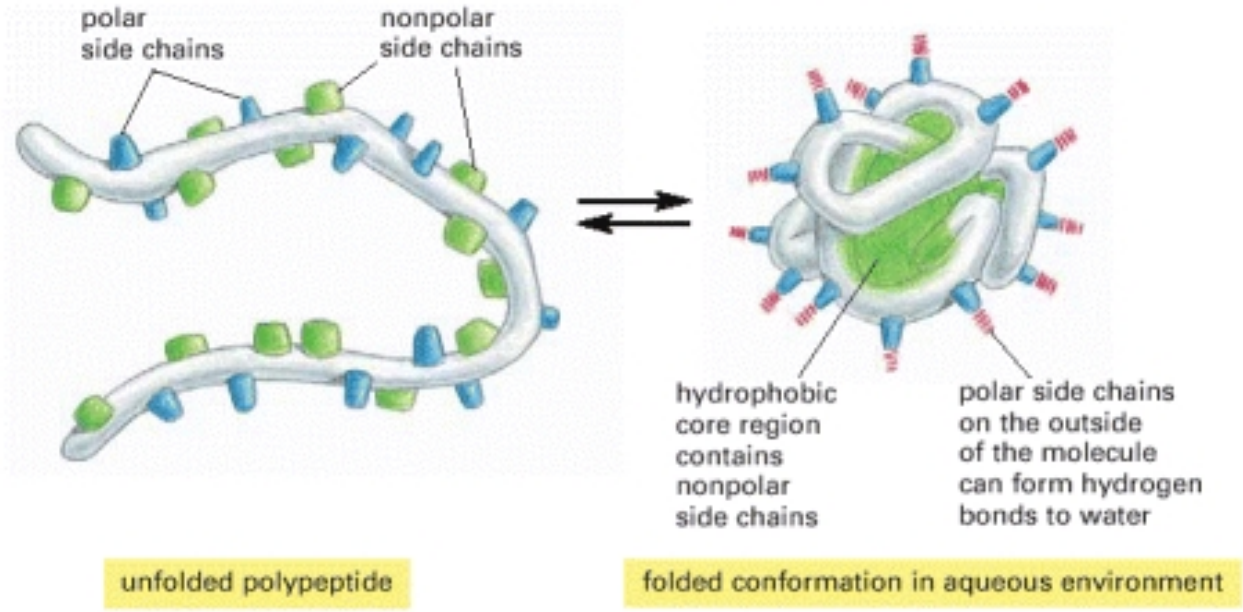
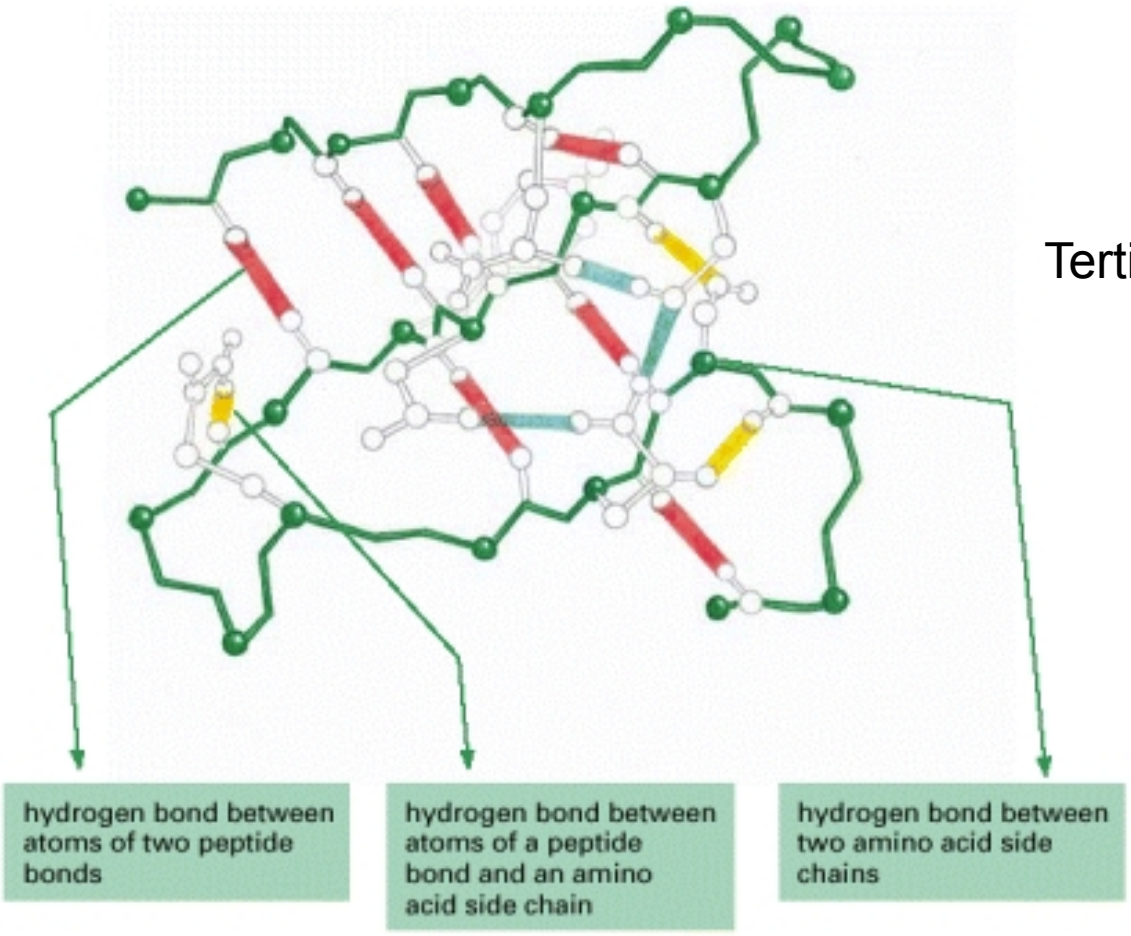


Image credit: Alberts et al., Mol Biol Cell

Side chains also form H-bonds



Tertiary structure

Image credit: Alberts et al., Mol Biol Cell

Reversible Unfolding of Tertiary Structure

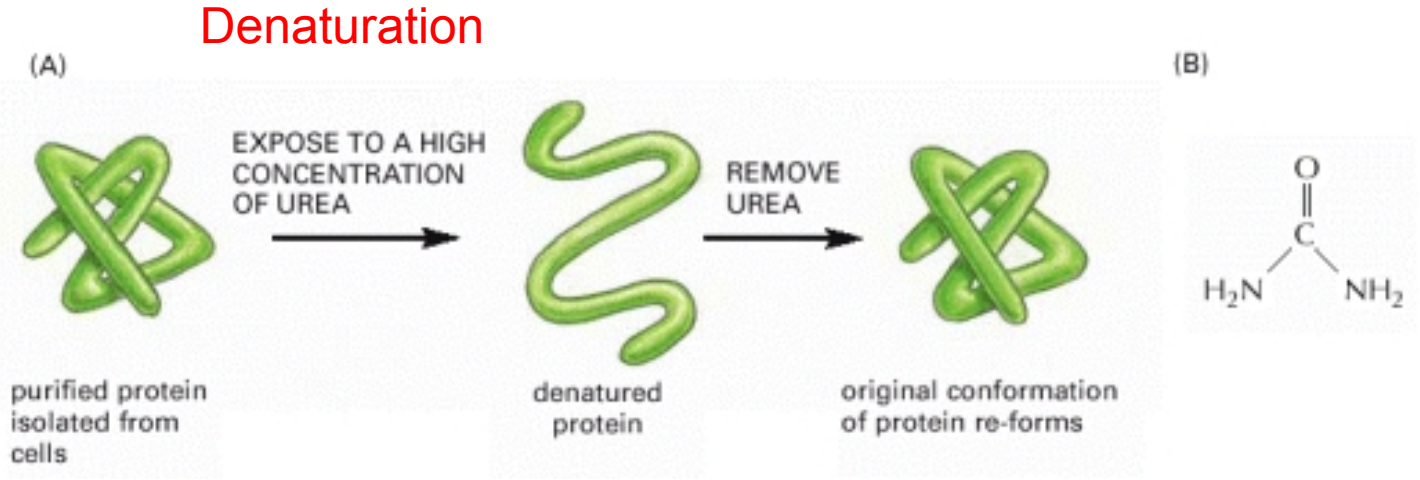


Image credit: Alberts et al., Mol Biol Cell

Quaternary Structure

Different Models of Src-ATP

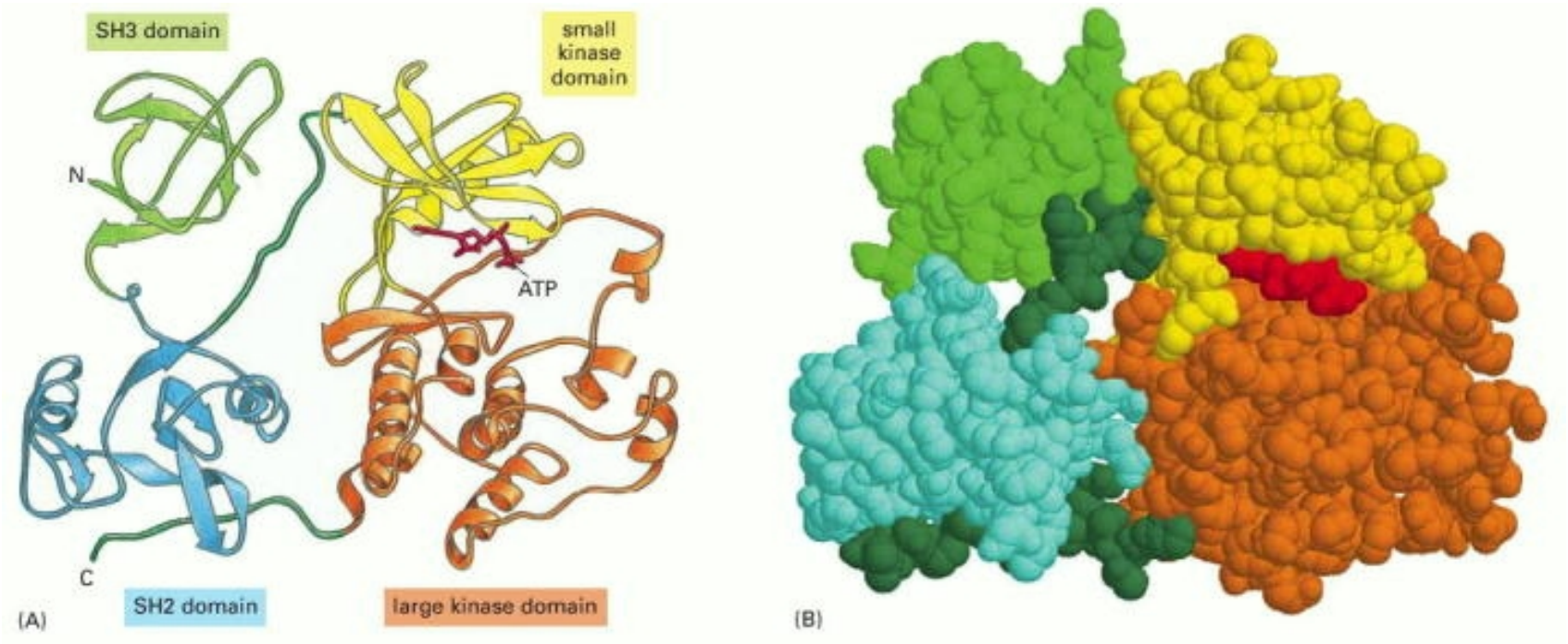


Image credit: Alberts et al., Mol Biol Cell

Common Globular Folds

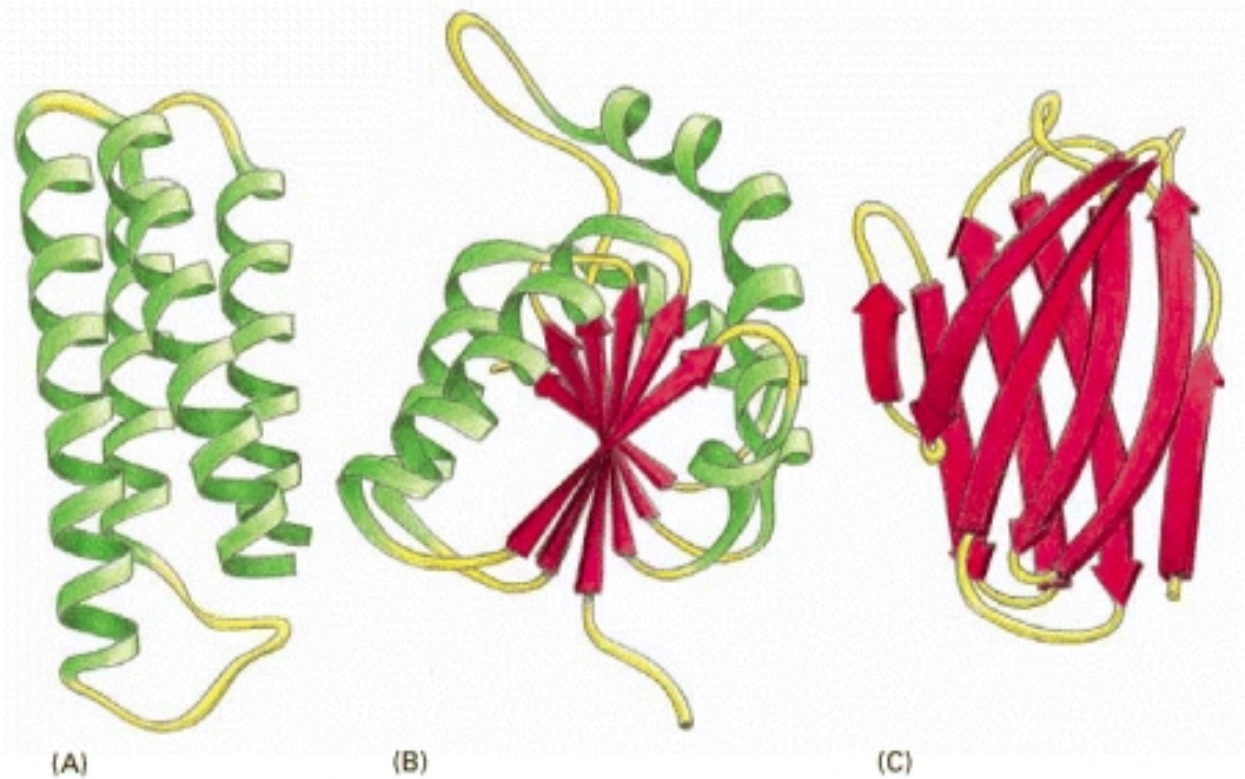


Image credit: Alberts et al., Mol Biol Cell

Compare two serine proteases

Green means sequence identity. Active site in purple.

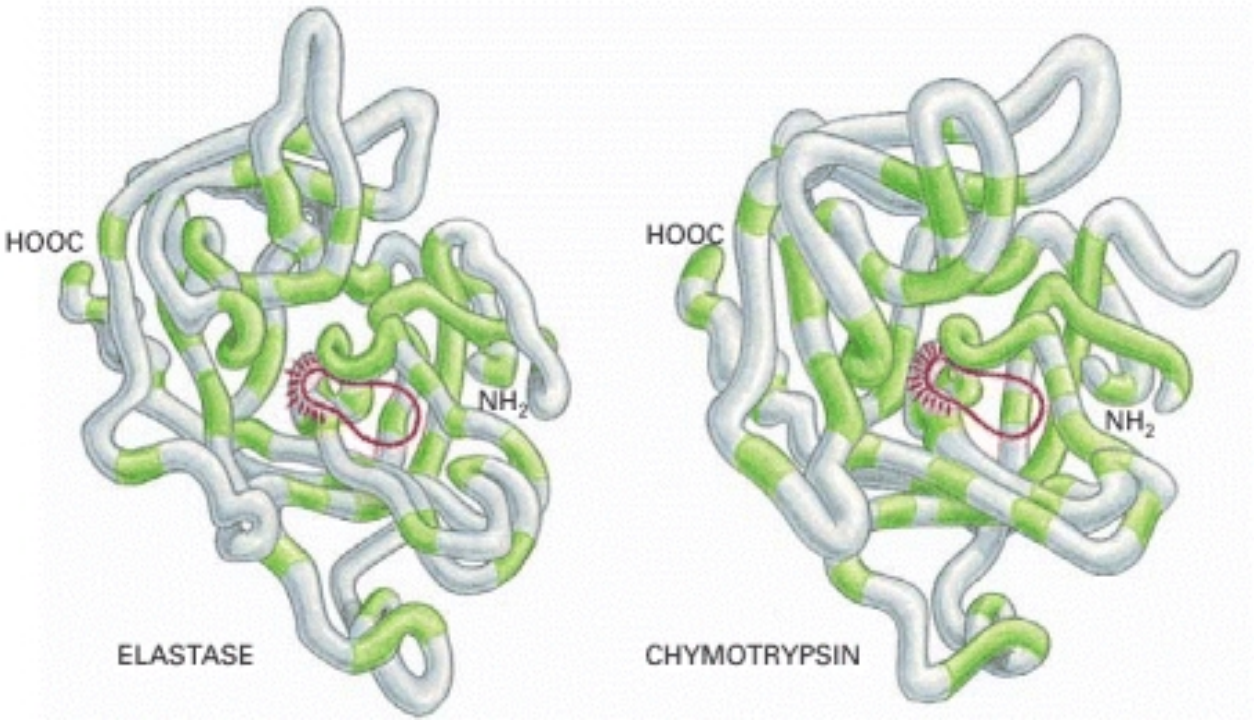


Image credit: Alberts et al., Mol Biol Cell

NUS National University of Singapore

Conserved structure in homeodomain proteins of yeast and Drosophila

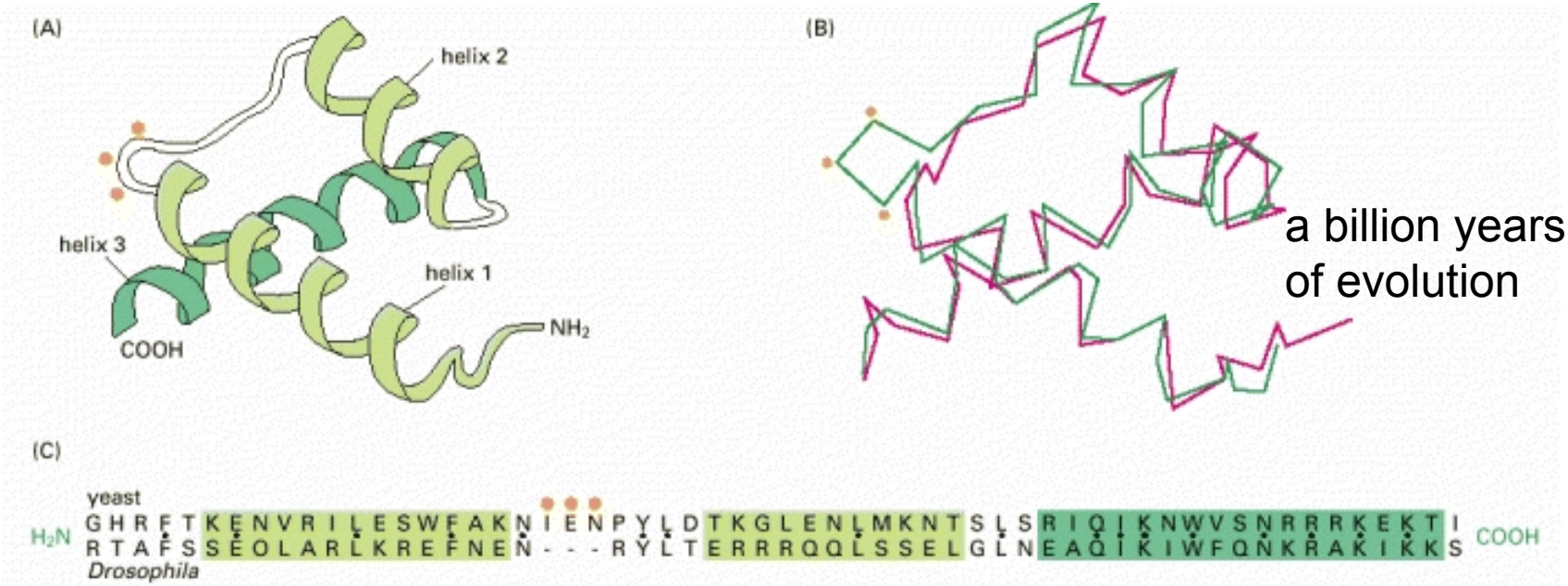
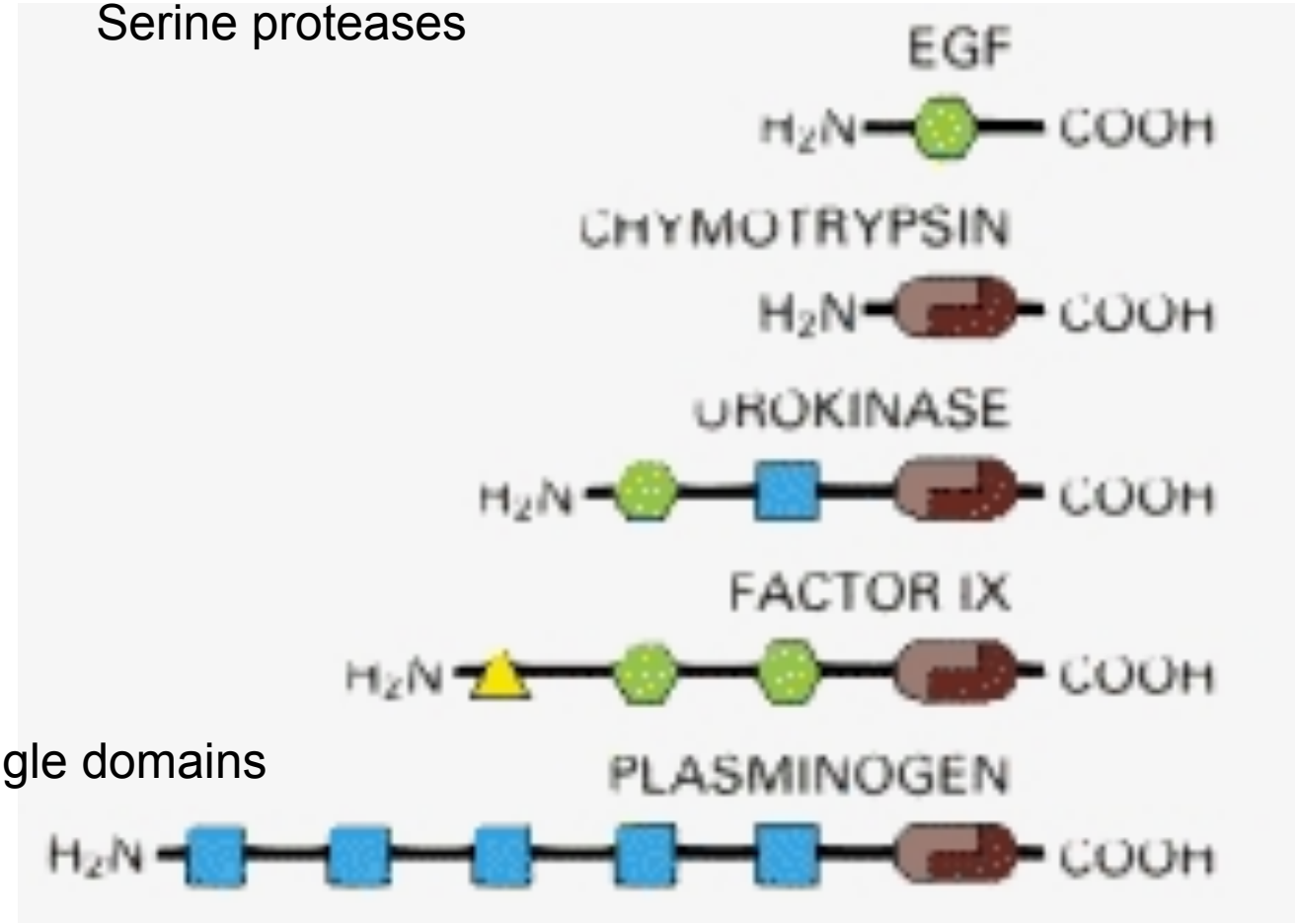


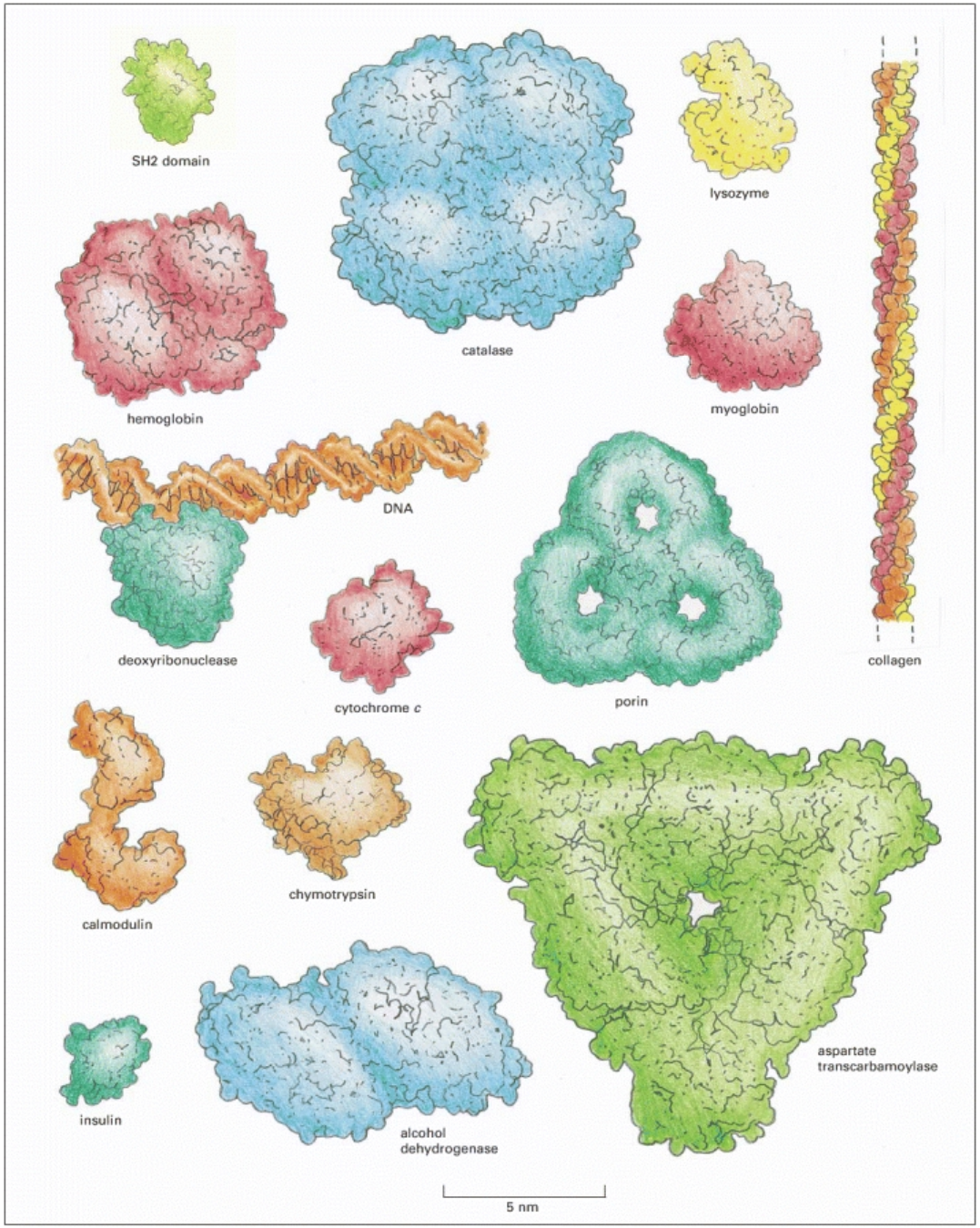
Image credit: Alberts et al., Mol Biol Cell

Domains Get Shuffled

Serine proteases



Kringle domains

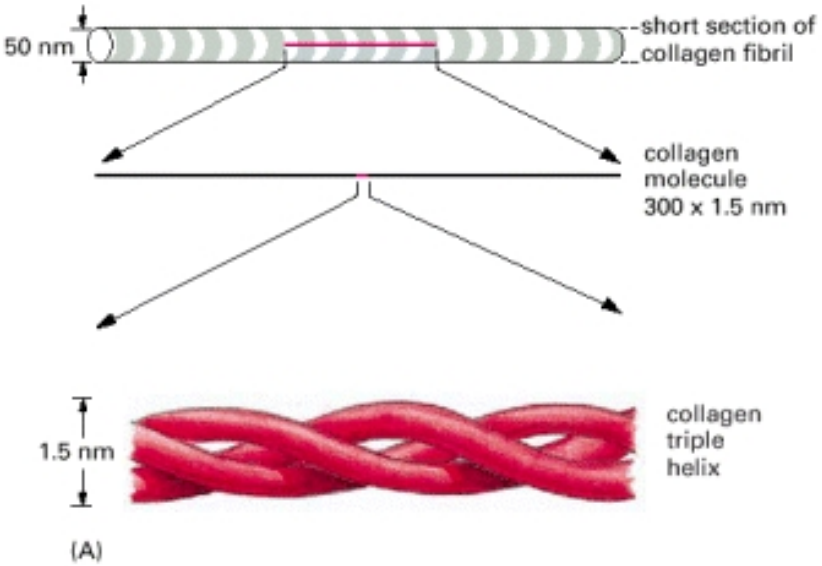


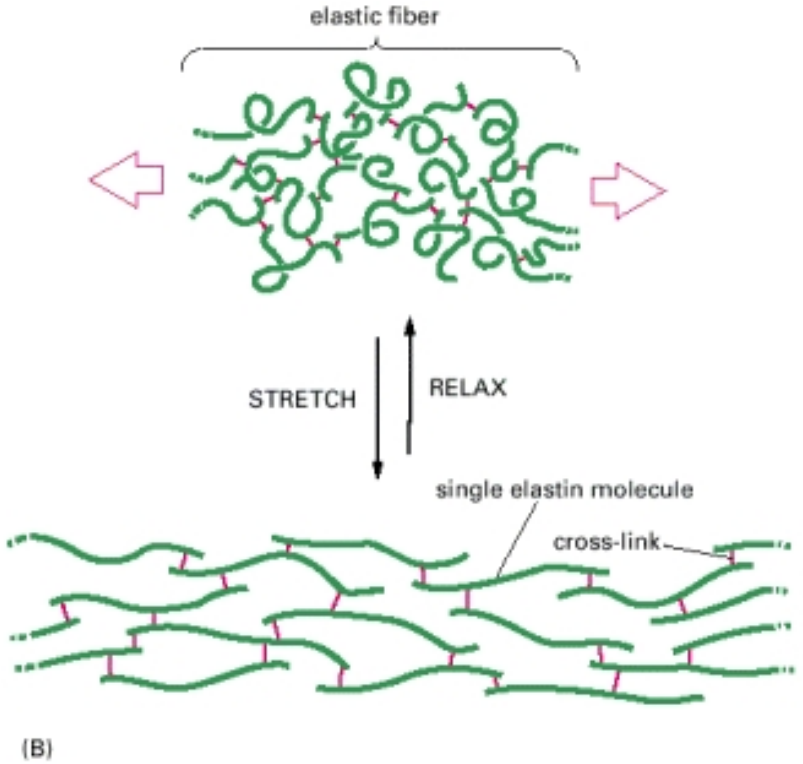
NUS National University of Singapore

Shown to scale

5nm = 50 Angstroms

Collagen and Elastin





Unextendable fibrils of collagen have tensile strength equiv to steel

Image credit: Alberts et al., Mol Biol Cell

NUS National University of Singapore

Ligands fit their binding pockets with high geometric complementarity, many weak interactions,

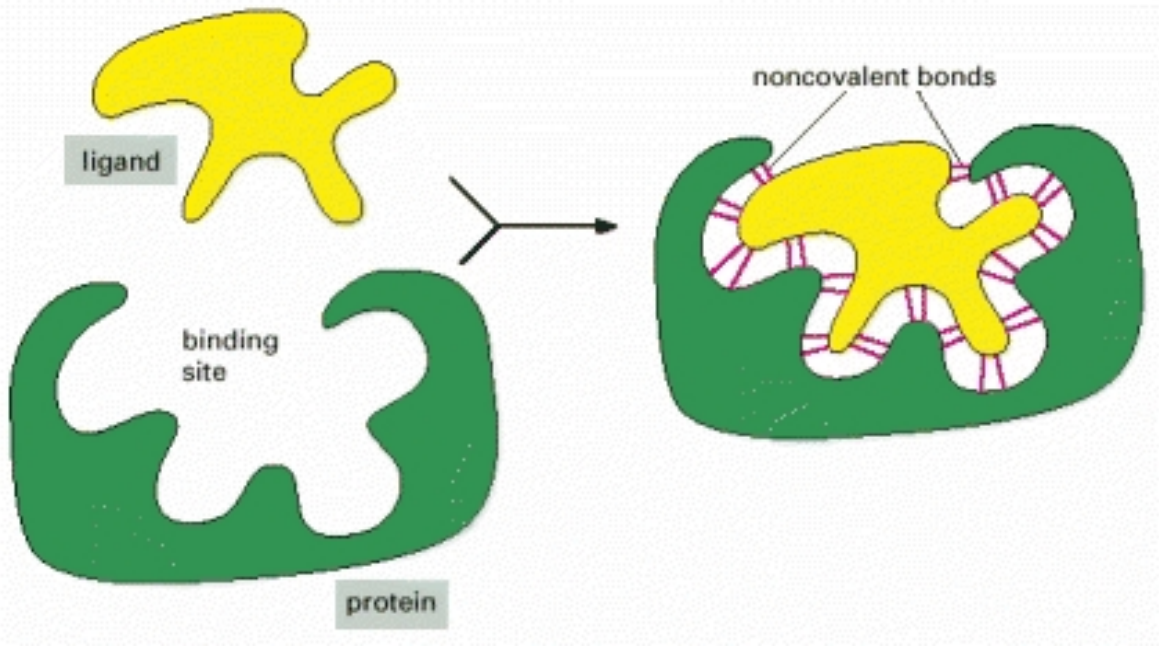


Image credit: Alberts et al., Mol Biol Cell

Random encounters are enough for specificity of interactions

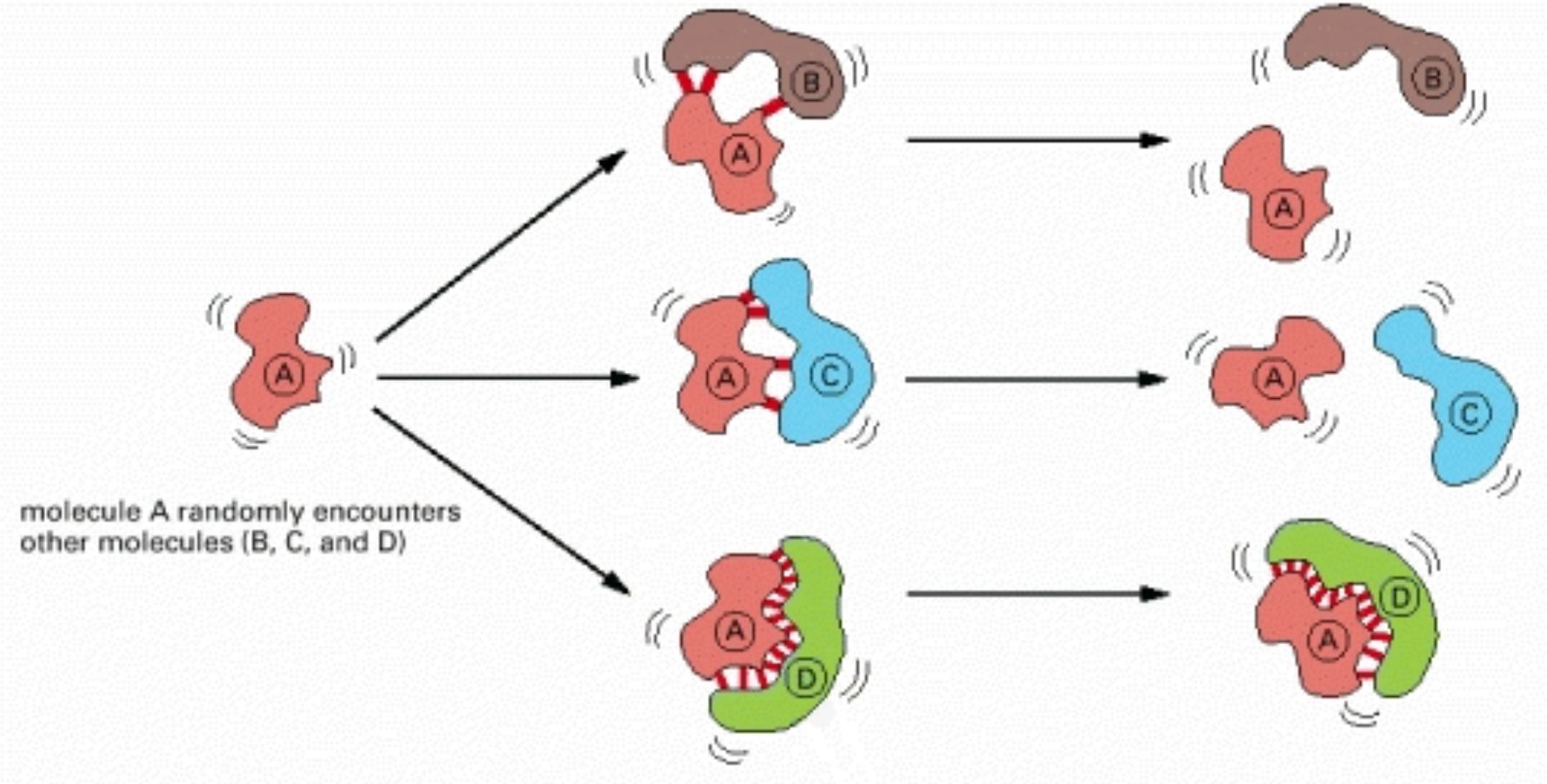


Image credit: Alberts et al., Mol Biol Cell

Cooperativity from Conformational Coupling:

NUS National University of Singapore

2 ligands drive the same conformational change, causing positive regulation

More interesting for computational modeling

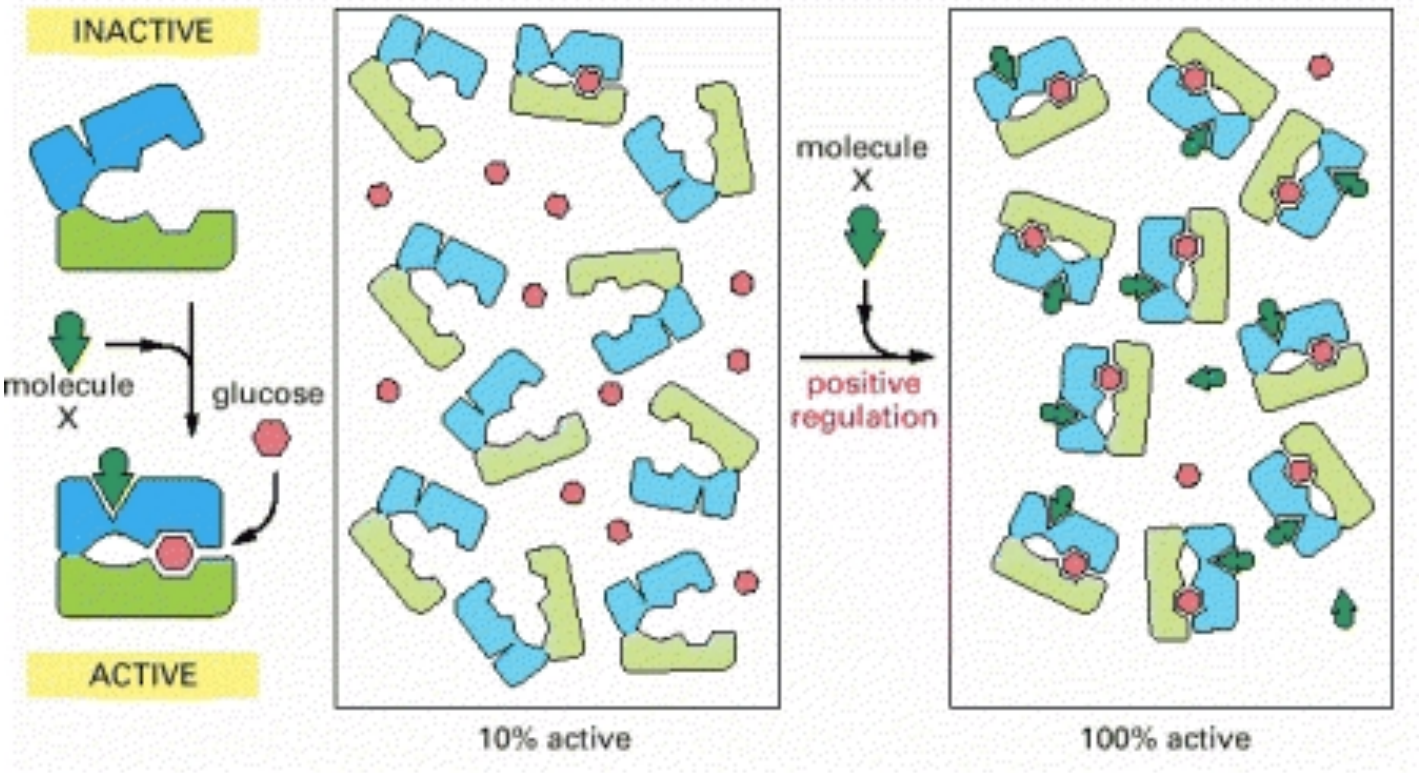


Image credit: Alberts et al., Mol Biol Cell

Protein functions can resemble logic gates

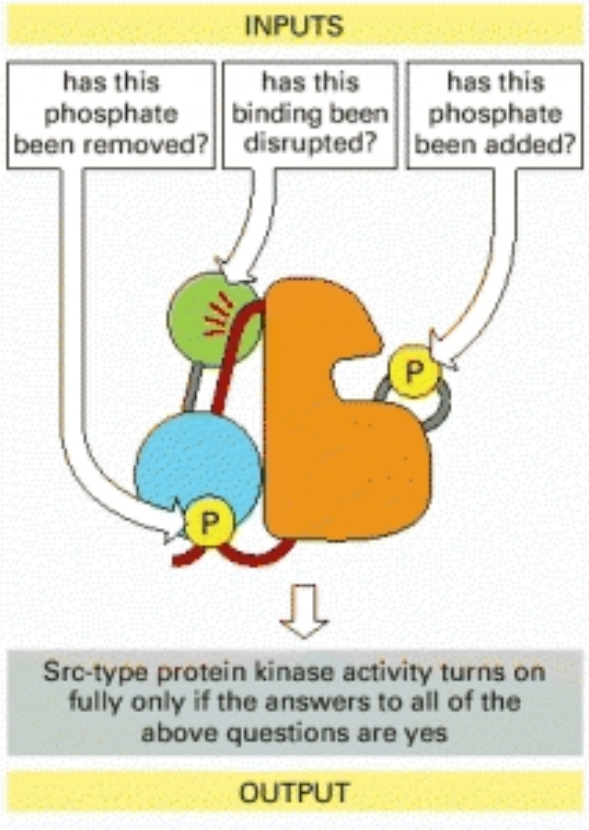


Image credit: Alberts et al., Mol Biol Cell

How biochemistry does it

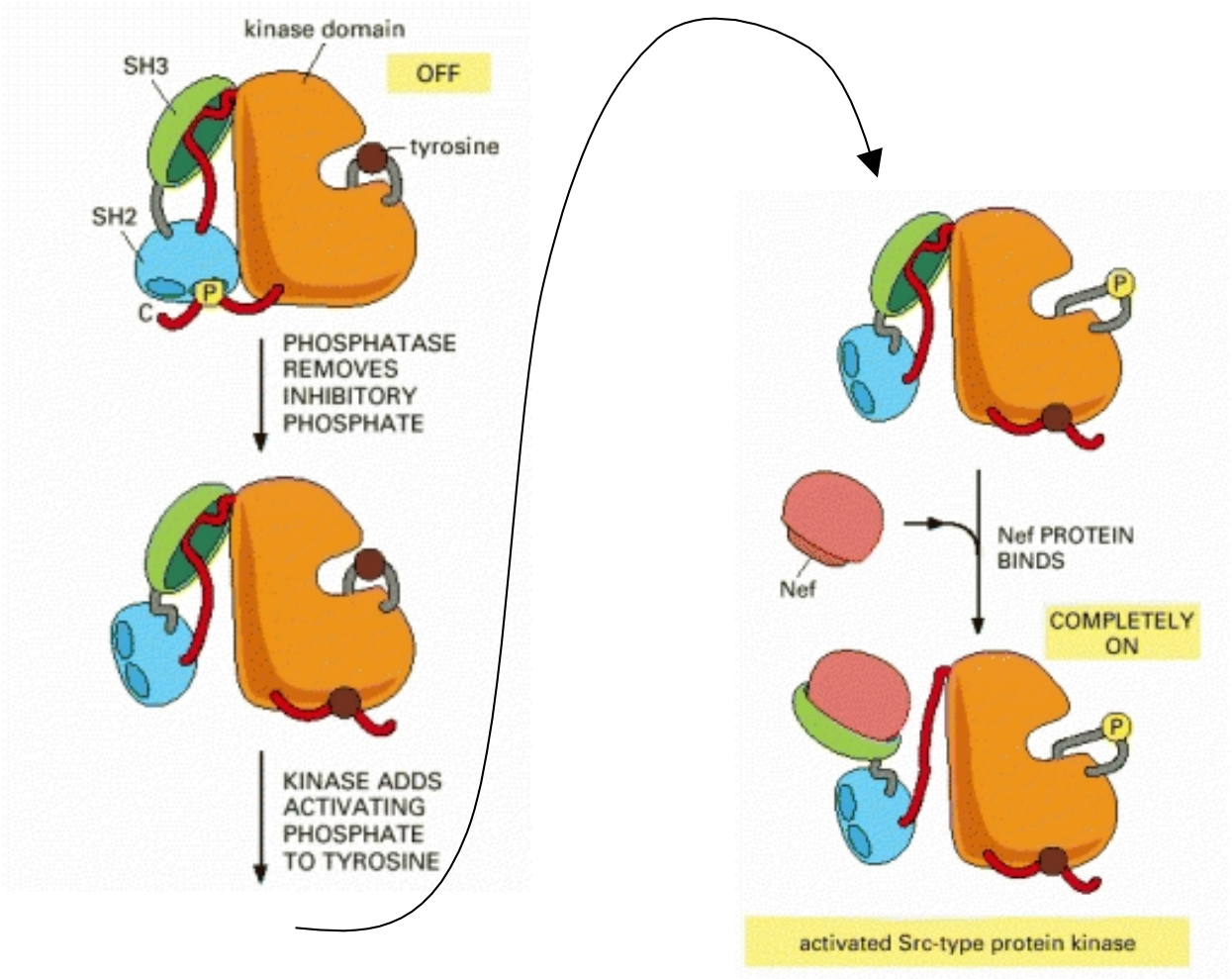


Image credit: Alberts et al., Mol Biol Cell

CDK kinase, another example

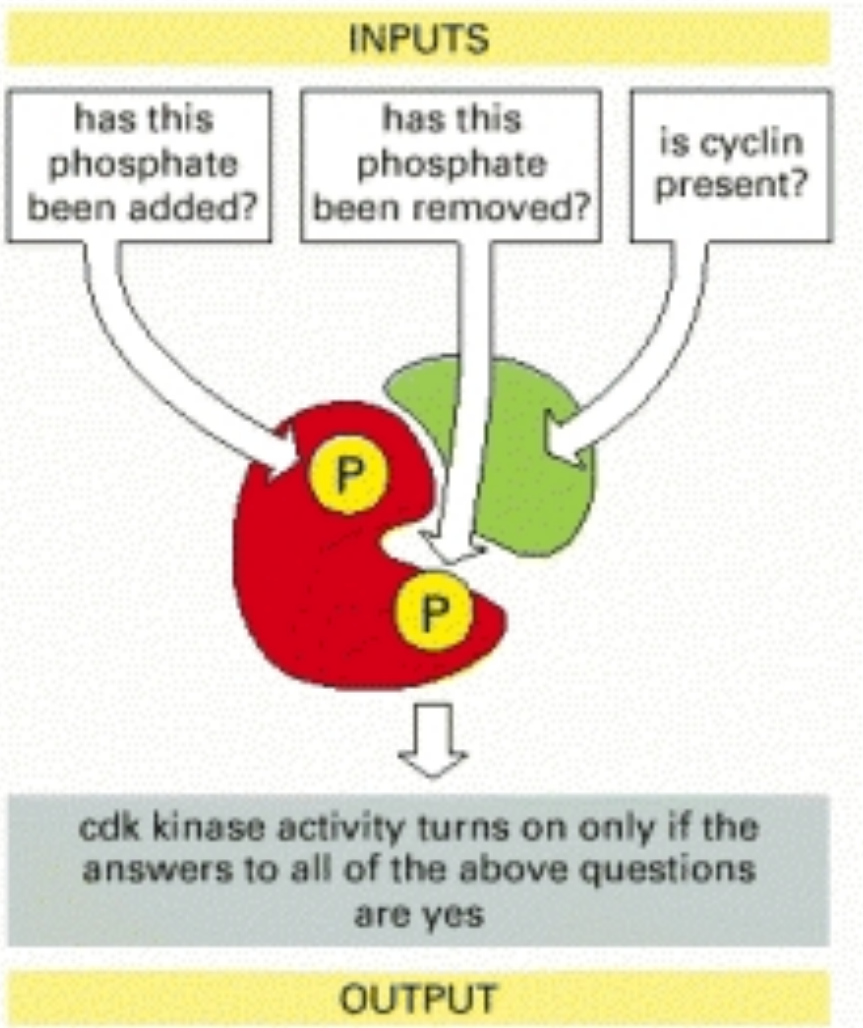


Image credit: Alberts et al., Mol Biol Cell

Motor Protein

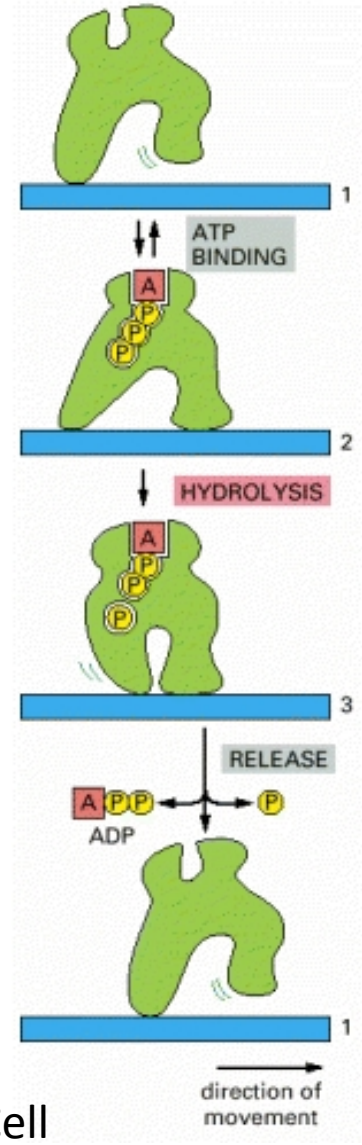


Image credit: Alberts et al., Mol Biol Cell

NUS National University of Singapore

Summary of Protein Structure and Protein Function

Structure

- **Primary**
- **Secondary**
- **Tertiary**
- **Quaternary**

Function

- **Specific Binding**
- **Allostery & Cooperativity**
- **Analogy with logic gates**
- **Example motor protein**

Molecular Modeling and Energy Functions

NUS National University of Singapore

Subsections

- **Molecular mechanics**
- **Potential Energy**
  - Electrostatic
  - Lennard-Jones
  - Dihedral, etc…
- **Simulating the physics**
  - Molecular Dynamics

NUS National University of Singapore

Molecular Mechanics

- **A molecule is described by interacting (soft) spheres.**
- **Different types of spheres describe different types of atoms.**
- **The interaction between atoms is described by special bonding or non-bonding interaction terms.**
- **The motion of all the atoms in the molecule is described by Newtonian classical mechanics.**

Use molecular mechanics to simulate a trajectory of the system: *Moving on an Energy Landscape*

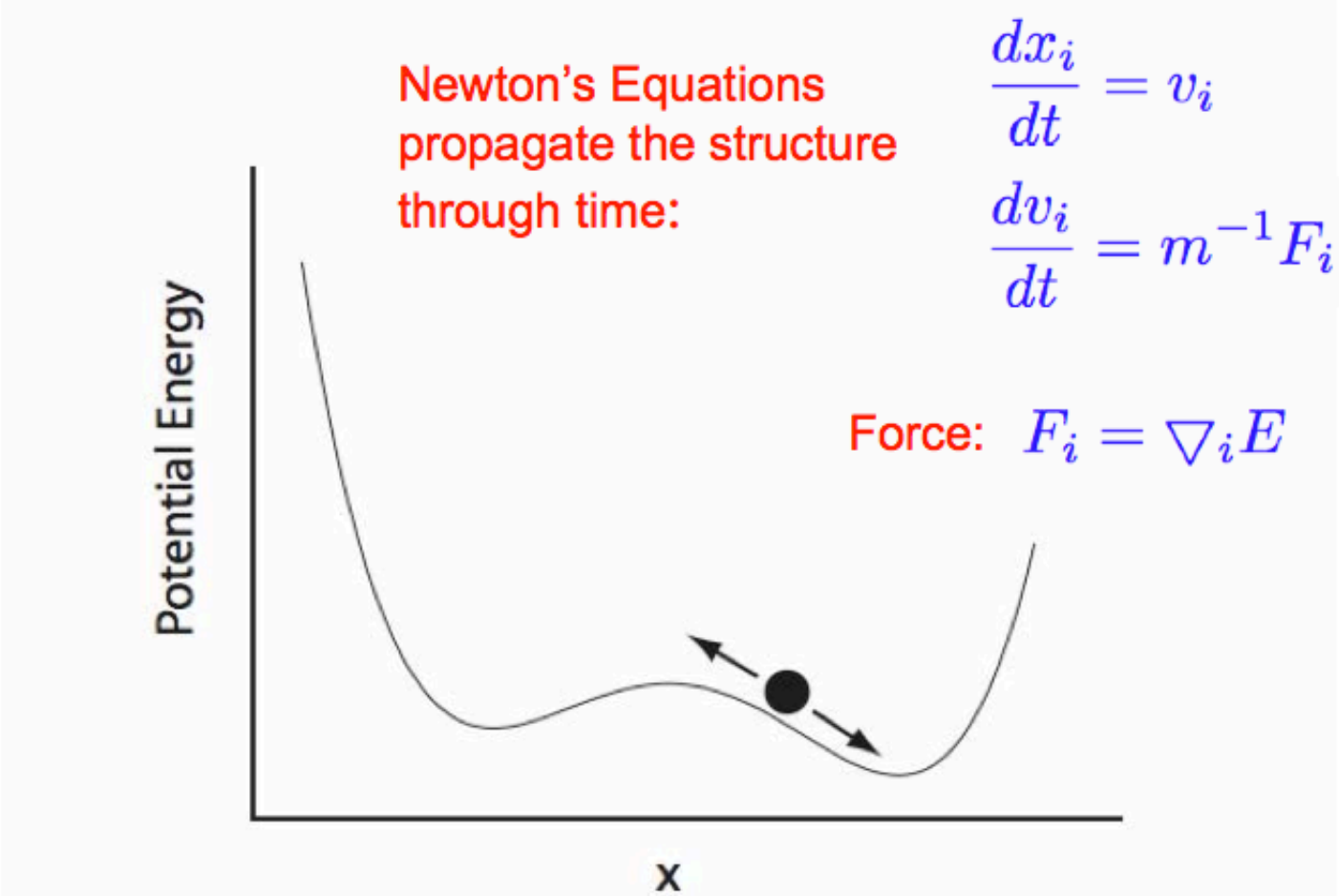


Image credit: John M. Robinson, biomachina.org

NUS National University of Singapore

Potential Energy

- **Don't confuse with Gibbs (ΔG = ΔH – TΔS)**
- **Potential energy** is energy stored in a system based on its conformation/configuration. (Stored in a molecule based on its conformation)
- This formula for an **electrostatic potential** should look familiar. (Epsilon is the dielectric constant)

$$W = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r},$$

- What about the potential energy when atoms overlap? Requires quantum chemstry?

NUS National University of Singapore

Potential Energy - Intuition

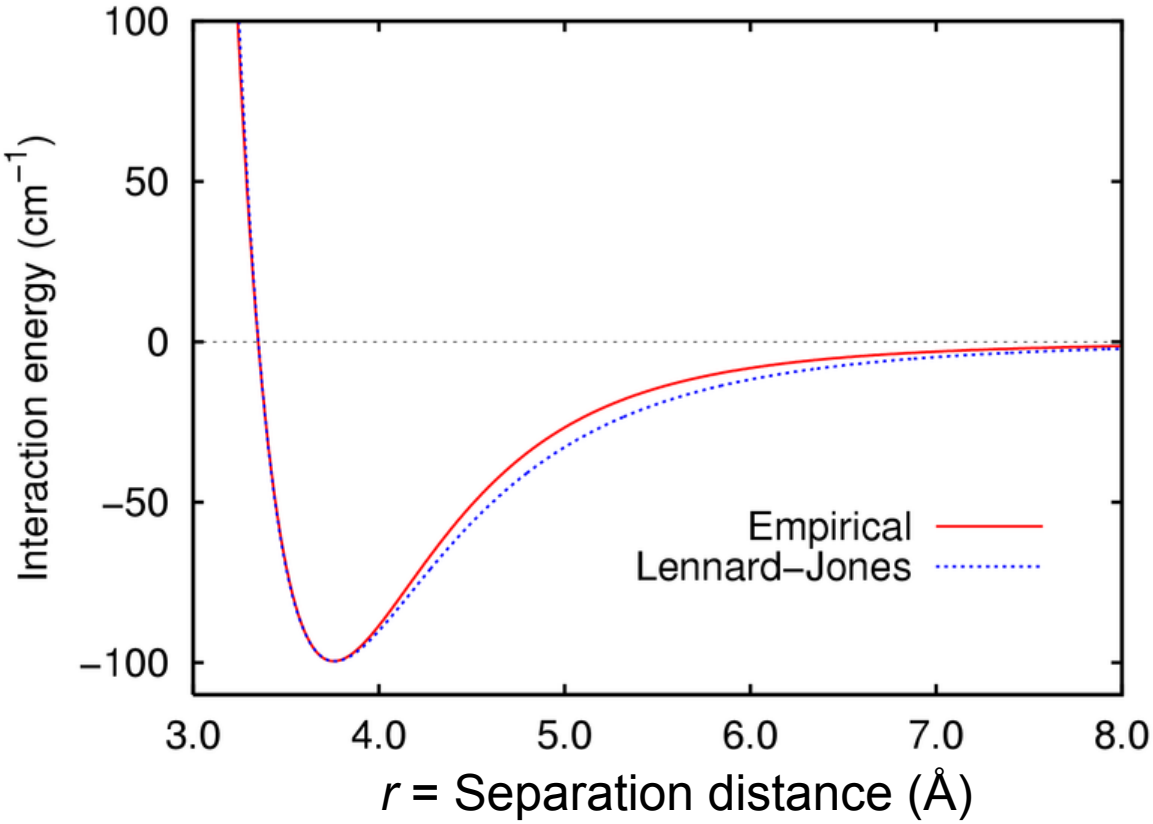
- **What does it mean if you're in a high energy state?**
  - It must have consumed energy to go from a lower energy state to where you are now.
  - There is a restoring force pulling you towards a lower energy state.
  - Real uses of potential energy involve differences between states rather than absolute levels.
- **What does it mean to have two charged particles with high electrostatic potential?**
- **What about a spring?**

$$W = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r},$$

Lennard-Jones Potential Energy

Simple example:
Attractive and repulsive forces between two molecules of argon.

Too close and they clash strongly
Too far apart and they have no energetic effect on each other.
What's in between?

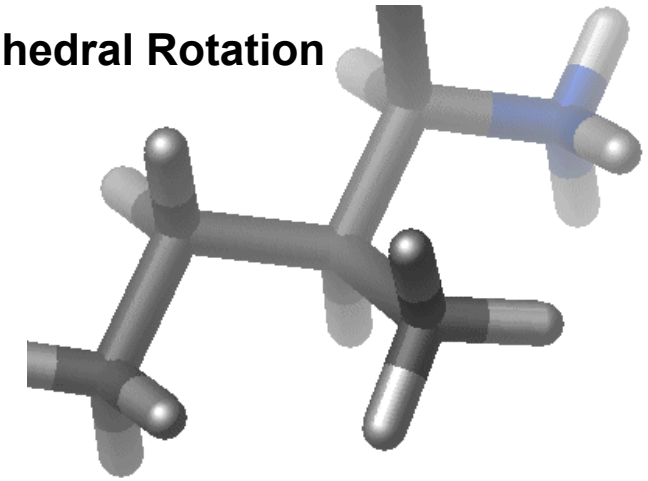


$$V_{LJ}(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right],$$

Protein Potential Energy

- $\vec{X}$ = 3-dimensional atomic coordinates for each atom in the protein

$$V_{total} = V_{electrostatic}(X) + V_{LJ}(X) + V_{dihedral}(X) + \ldots$$



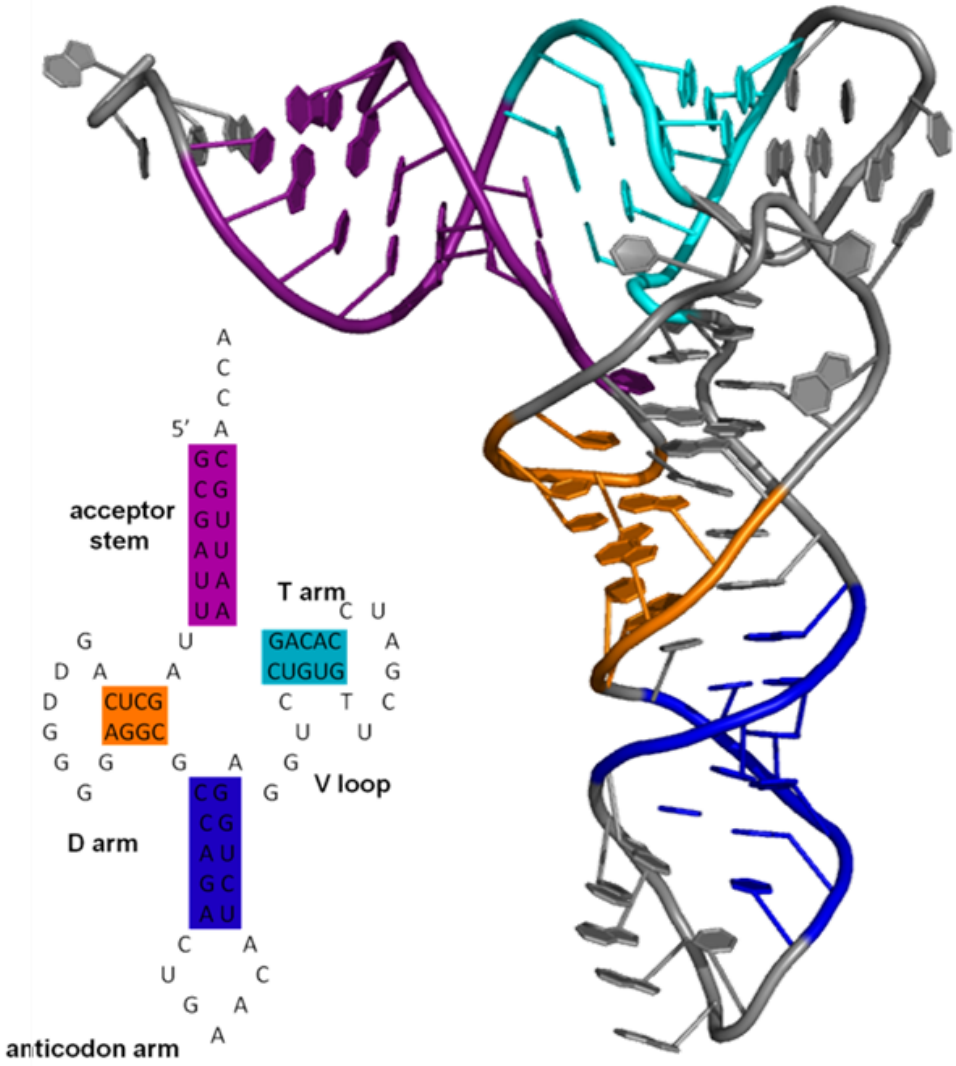
Other potentials typically used for proteins: covalent bond lengths and angles, dihedral angles, improper angles.

Movie from CHARMM tutorial at embnet.org

NUS National University of Singapore

RNA Potential Energy

- **Exercise: Try extending the protein example but for a different type of polymer.**
- **V_{tot} = ? + ? + …**



NUS National University of Singapore

Water

- **Protein configurations are highly dependent on the surrounding solvent, generally mostly water.**
- **Electrostatic effects can be massively screened by solvation.**
- **Water is trickier to model than it looks. Many of the issues are entropic, which we're not dealing with in just the potential energy. (not Gibbs)**
- **There are some widely different approaches for dealing with water**
    - ***Implicit: Distance-dependent dielectric***
    - ***Implicit: Poisson-Boltzmann***
    - ***Explicit: Solvent***

Molecular Dynamics Simulations

1. **Initialize the protein in a configuration.**
2. **Use the potential energy to compute the force vector on each atom.**
3. **Move the atoms according to those forces**
   - For an infinitesimal length of time, we can ignore changes in the force. Fixed force means we can move the atoms linearly. (Computationally easy)

   (or is it easy?)
4. **Return to step 2 until sufficient duration has been simulated**

Molecular Dynamics continued

- **The reason it's not easy is because the infinitesimal lengths of time have to be really short for any decent results**
  - Femtosecond (10^{-15} second) sized steps

- **Setting up MD simulations is trickier than it looks. After that, performing the simulations is easier than it looks, except for consuming time.**
- **How do you know you did it properly?**
- **What do you do with it afterwards?**

NUS National University of Singapore

Why do simulations?

- **What are the functional motions?**
  - Molecular (deterministic) dynamics
  - Stochastic dynamics
  - Normal modes

Dynamics

- **What are the most probable conformations?**
  - Molecular dynamics
  - Monte Carlo methods
  - Hybrid MD-MC methods

Sampling

- **What is the most stable (probable) structure?**
  - Energy minimization
  - Simulated annealing

Optimization

Optimization

NUS National University of Singapore

Optimization

- **Minimizes or maximizes an objective function with respect to decision variables.**
  - May also involve constraints
- **A.k.a. "Mathematical programming"**
- **The term optimization is used in other contexts, such as database optimization.**

Any past experience with optimization?

NUS National University of Singapore

Using Calculus

You probably remember finding the roots of a function in calculus class.

How is that helpful for optimization?

NUS National University of Singapore

Optimization Approaches

- **Deterministic Global**
  - e.g., solving roots
  - Not all functions have tractable deterministic global optimization methods, but finding ways to make your problem fit tools of this sort is very important. (not the topic we address today)
- **Local**
  - e.g., gradient descent
- **Global but stochastic and/or heuristic**
  - e.g., simulated annealing, genetic algorithms

Any questions?

NUS National University of Singapore

Local Descent

- **S is the search space of possible solutions.**
  - "feasible"
- **f is the objective function mapping S to the real numbers.**
  - "score"
- **s_0 is the initial state, and $N(s_i)$ are the neighbors of a state s_i**
- **Now let's figure out pseudo-code for the rest…**

```
Initialize …
      Loop ….
Terminate …
```

NUS National University of Singapore

Local Descent, continued

- $N(s_i)$ is usually defined by simple potential moves, e.g. swapping some elements of the current solution by non-solution elements.
- If k elements are swapped then we have k-neighborhoods and we obtain k-optimal solutions.
- Steepest descent searches all solutions in $N(s_i)$ and chooses the best improving one.
- Random descent selects neighboring solutions randomly and accepts the first improving one
- Final solution is dependent on starting point. Why?

Guess what the hill-climbing method does.

Evolutionary Strategies & Genetic Algorithms

- **Algorithms inspired by an analogy between optimization and evolution.**
- **Setting up the problem, how can we translate ideas of evolutionary biology into an algorithm?**
  - Population of organisms
  - Mating
  - Time
  - Inheritance
  - Mutation
  - Selection

Genetic algorithms tend to be less popular for protein structure applications than another randomized heuristic algorithm from the same era…

NUS National University of Singapore

Simulated Annealing

- **A greedy method that only goes downhill gets stuck in local minima. But if you are allowed to go uphill, how do you avoid wandering aimlessly and wasting all your time in high-energy states?**
- **Based on physical analogy: how do physical systems arrive at an energetically optimal state?**
  - Real systems can get stuck in non-optimal states, e.g. sudden cooling into a glass, not a crystal.
  - Get a stuck system unstuck by heating and the slow cooling.

"The method was independently described by Scott Kirkpatrick, C. Daniel Gelatt and Mario P. Vecchi in 1983, and by Vlado Černý in 1985. The method is an adaptation of the Metropolis-Hastings algorithm, a Monte Carlo method to generate sample states of a thermodynamic system, invented by N. Metropolis et al. in 1953."--Wikipedia

NUS National University of Singapore

Simulated Annealing

A randomized heuristic method for global optimization.

- **Each step of the SA algorithm considers a random "nearby" state and then shifts to that state with a probability that depends on the energy difference and on a global temperature parameter *T*.**

Have you learned about the Boltzmann distribution?

$$\frac{N_i}{N} = \frac{g_i e^{-E_i/(k_B T)}}{Z(T)}$$

- **In statistical thermodynamics, the probability of an increase in energy ΔE is**

  exp(-ΔE/kT) where T=temp, k=Boltzmann's constant

(Simulated annealing doesn't require using this choice of distribution)

NUS National University of Singapore

Simulated Annealing

- **The global temperature is high during early steps and gradually lowered later.**
  - The “annealing schedule” is a choice to be made when setting up the algorithm.
- **If using the Boltzmann probability distribution,**
  - at high temperature, the movements are essentially random.
  - at zero temperature the movements are strictly downhill.
  - **What are the movements at intermediate T?**

NUS National University of Singapore

Using Simulated Annealing for Protein Structure Optimization

- **Setting up the problem**
  - What are the decision variables?

    (x,y,z) coordinate positions of all the atoms
  - What is the objective function?

    The potential energy function
  - What constitutes a random step of change?

    This is an interesting question. Remember protein structure?
  - What is the probability of accepting a change?
  - What is the annealing schedule?
- **What will be the output of running SA?**

General guidelines for annealing

Source: Nur Evin Özdemirel and *Modern Heuristic Techniques for Combinatorial Problems*, Wiley, 1993

- **Choose the annealing schedule so the high temperature will cause the proportion of accepted moves (both improving and non-improving) to reach a prespecified value**

- **For each new temperature, allow the system to reach nearly steady state before lowering the temperature again.**

NUS National University of Singapore

General guidelines for what is a step

Source: Nur Evin Özdemirel and *Modern Heuristic Techniques for Combinatorial Problems*, Wiley, 1993

- **Neighborhood structure is problem-specific but at the least every state should be indirectly reachable from every other state (for convergence and for validity)**

- **Random generation of a neighboring solution should be fast for effective use of computation time, but generating a feasible solution can be tricky.**
  - An ideal neighborhood structure also facilitates a rapid computation of ΔV.

Optimizing Pseudo-Energy

Pseudo-Potential Energy

Replace or augment the real potential energy with some other function

Why would we do this?

Image credit: Arkady Krasheninnikov, University of Helsinki

Why optimize a Pseudo-Energy instead of the real Potential?

- **Decrease harsh penalties and remove barriers**
  - To smooth the gradients so the landscape is better-conditioned for the optimizer.
  - Allows more efficient minimization and fewer local minima.
- **Push towards a desired destination**
  - E.g., for making a protein adjust structure to allow docking with a ligand
- **For testing plausibility of a suspected state**
  - E.g., trying a homologous template structure
- **Force the structure to agree with experimental evidence**
  - E.g., for determining the structure.

Why optimize a Pseudo-Energy instead of the real Potential?

NUS National University of Singapore

- **Decrease harsh penalties and remove barriers**
  - To smooth the gradients so the landscape is better-conditioned for the optimizer.
  - Allows more efficient minimization and fewer local minima.
- **Push towards a desired destination**
  - E.g., for making a protein adjust structure to allow docking with a ligand
- **For testing plausibility of a suspected state**
  - E.g., trying a homologous template structure
- **Force the structure to agree with experimental evidence**
  - E.g., for determining the structure.

Important

NUS National University of Singapore

Structure Determination

- **Atomic-level information (< nm) can only be probed indirectly**
  - X-ray diffraction from a crystal
  - Transfer magnetization between nearby dipoles
- **Experimental molecular biophysics produces very powerful datasets that can be a nightmare to interpret in terms of atomic coordinates**
  - Intensities but not phases of a Fourier transform of the electron clouds.
  - Distance restraints (through space) or dihedral restraints (through bonds).

How to make sense of indirect partial clues?

NUS National University of Singapore

Example: NMR Spectroscopy

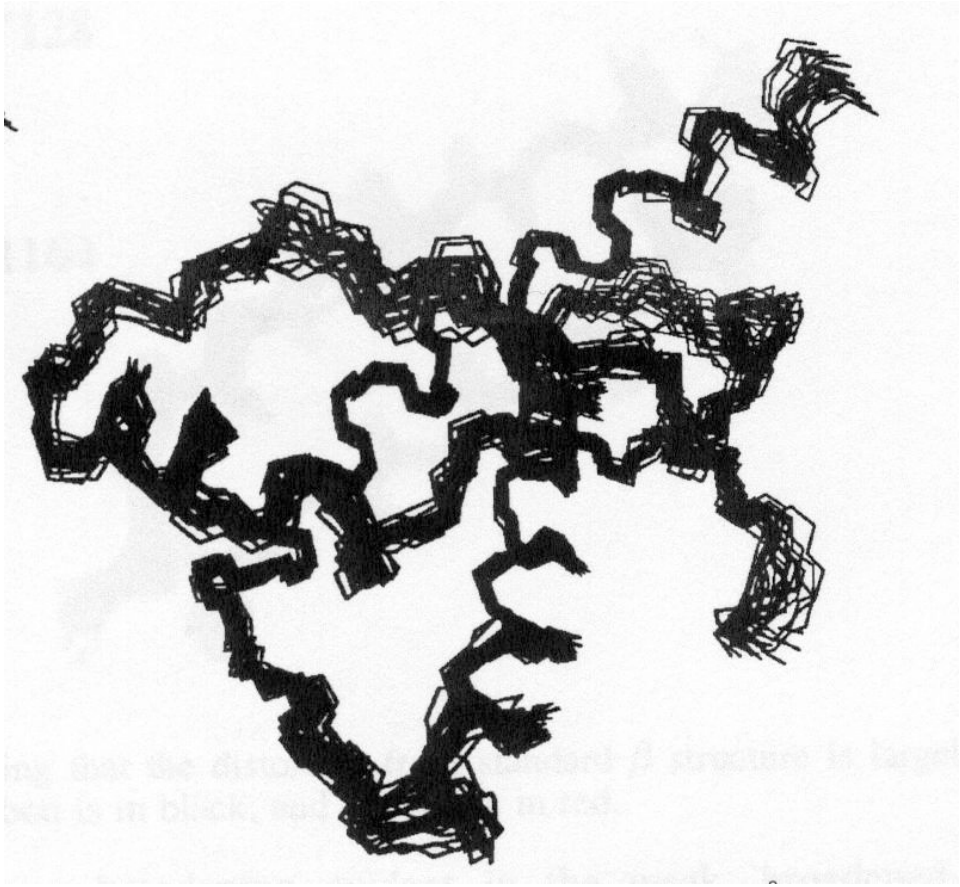
Source: Matt Cordes

- Method: The sample is prepared, resonances are assigned to the nuclei, restraints are generated and a structure is calculated from the restraints.
- An ”NOE” (crosspeak in NOESY experiment) signifies spatial proximity between the two nuclei in question. Each such observation can be converted in to a maximum distance between the nuclei, usually between 1.8 and 6 Ångstroms.
- Knowing the distances between all atoms would be sufficient to determine the whole structure.
- Can we determine the structure using many observations of approximate atomic distances?

Example: NMR spectroscopy

Source: Matt Cordes

- Background Concept: Ensemble of structures
- NMR structures are reported as ensembles, giving them a “fuzzy” appearance.
  - This is informative and sometimes annoying



Example Pseudo-potential for NMR

Source: Matt Cordes

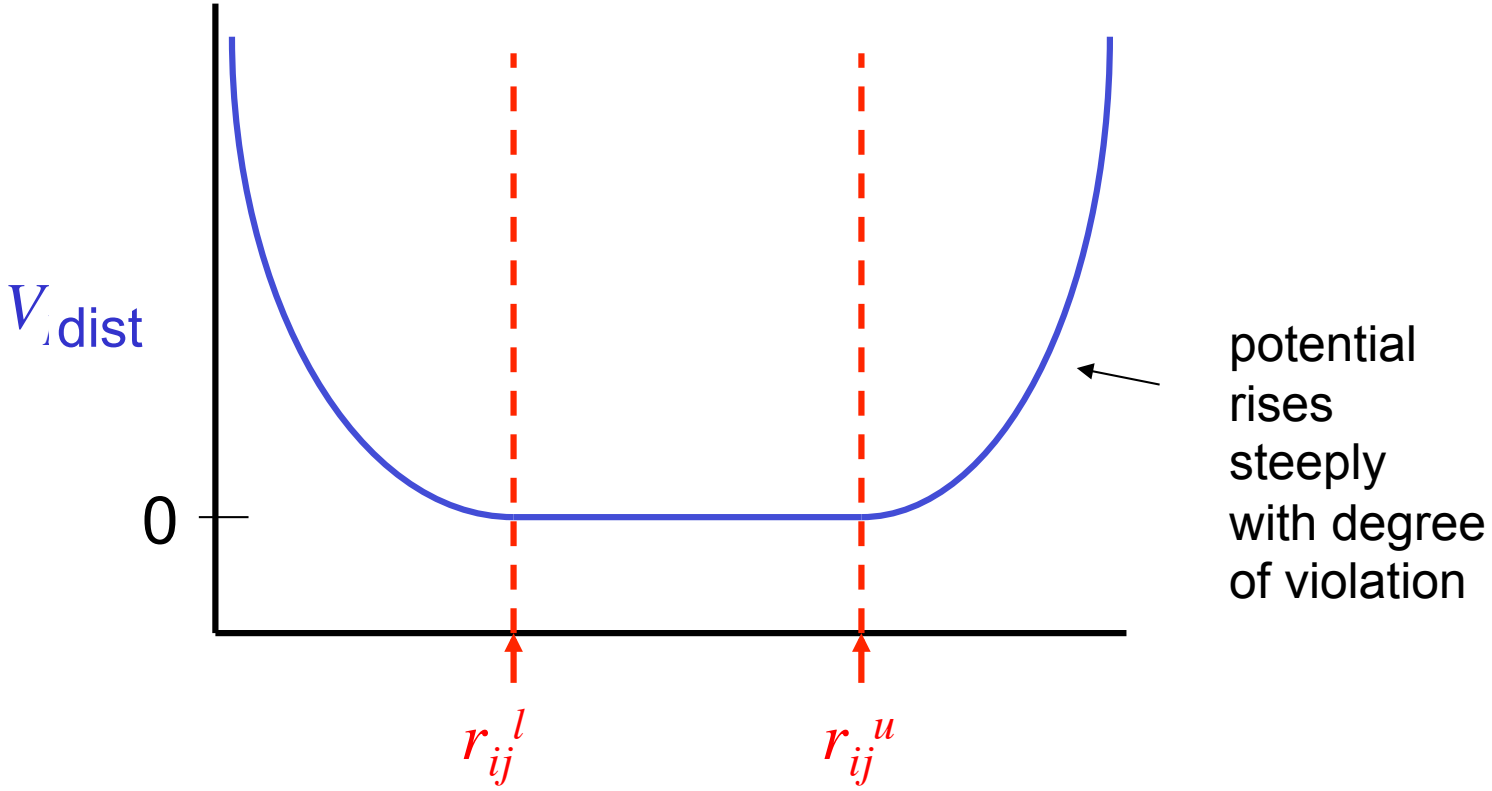
- **Generate fake energy amounts that represent the cost of violating the experimental constraints on distances or angles. Turn the constraints into "restraints."**
- $V_{total} = V_{bond} + V_{angle} + V_{dihedr} + V_{vdW} + V_{coulomb} + V_{NMR}$

$$V_{dist} = \begin{cases} k\,(r_{ij}-r_{ij}^{U})^2 & \text{if } r_{ij} > r_{ij}^{U} \\ 0 & \text{if } r_{ij}^{L} < r_{ij} < r_{ij}^{U} \\ k\,(r_{ij}-r_{ij}^{L})^2 & \text{if } r_{ij} < r_{ij}^{L} \end{cases}$$

- where r_{ij}^{L} and r_{ij}^{U} are the lower/upper bounds and k is a force constant (~ 250 kcal mol^{-1} nm^{-2}). This makes it somewhat permissible to violate restraints by paying an energetic penalty.

Example pseudo-potential

Source: Matt Cordes



r_{ij}^{L} and r_{ij}^{U} are the lower/upper bounds

Example of Using Optimization: Solving Atomic Structures by NMR

Source: Matt Cordes

After we run the optimizer are we done?

- It's typical to generate 50 or more trial structures, but not all will converge to a final structure that is physically reasonable or consistent with the data.
  - Must discard some from inclusion in the ensemble
- Some typical *acceptance criteria*:
  - no more than 1 NOE distance restraint violation greater than 0.4 Å
  - no dihedral angle restraint violations greater than 5 degrees
  - no gross violations of reasonable molecular geometry
- Some possible rejection criteria:
  - Too many residues with backbone angles in disfavored regions of Ramachandran space
  - Too high a final potential energy
- The ensemble and minimized ave get published.

Example of Using Optimization: Solving Atomic Structures by NMR

- **What are the take-home messages?**
- **I don't expect you to understand NMR**
- **See an example of how optimization is used in the field of computational structural biology**
- **Prime your imagination to see other opportunities to use**

Concluding Remarks

NUS National University of Singapore

What have we learned?

- **Protein Structure and Function**
  - Primary, secondary, etc. levels of structure
  - The potential energy as a mathematical representation of the structural rules
- **Brief Introduction to Optimization**
  - Example of simulated annealing as an optimization method
- **Optimizing pseudo-energy functions is a popular way to solve many problems in structural biology**
  - Example of interpreting NMR data

Any Questions?

NUS National University of Singapore

Acknowledgements

- **Thanks to Matt Cordes, Univ Arizona, Biochemistry Dept.**
  - Lecture on computational analysis of NMR from his course on Biomolecular Structure

- **Molecular Biology of the Cell by Alberts et al., is on the internet, available to the public, with the creators properly compensated, thanks to the US's National Library of Medicine (NCBI).**

NUS National University of Singapore

References

- **Potential energy:** http://www.ch.embnet.org/MD_tutorial/pages/MD.Part2.html
- **Optimization:** The field is broad and I don't know of any good textbooks for the breadth of optimization methods, but there are books on super-families like deterministic nonlinear optimization methods, stochastic optimization heuristics, etc.
- **NMR spectroscopy:** "De novo determination of peptide structure with solid-state magic-angle spinning NMR spectroscopy." By by Rienstra, Tucker-Kellogg, Jaroniec, Hohwy, Tidor, Lozano-Pérez, Griffin. PROC. NATL. ACAD. SCI. USA, 2002 Aug 6: 99(16): 10260-5.
- **X-ray crystallography:** "Engrailed Gln50 - Lys homeodomain-DNA complex at 1.9 A resolution: structural basis for enhanced affinity and altered specificity" by Tucker-Kellogg, Rould, Chambers, Ades, Sauer, and Pabo. STRUCTURE, Vol. 5, No. 8, pp. 1047-1054, 1997. PMID: 9309220