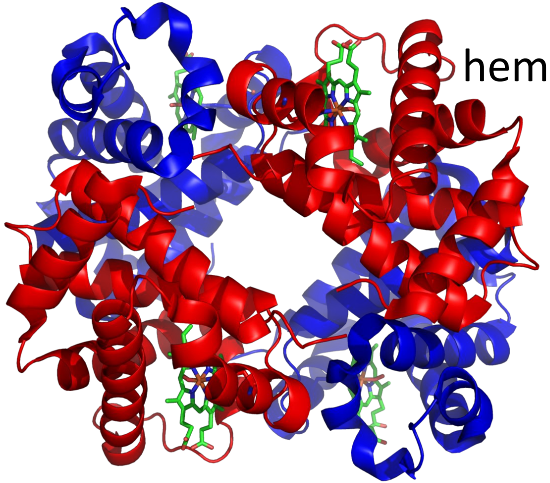
UMSL | Department of Mathematics & Computer Science

Simulated Annealing for Protein Folding

Proteins We Eat & Proteins Inside Our Body

Proteins we eat

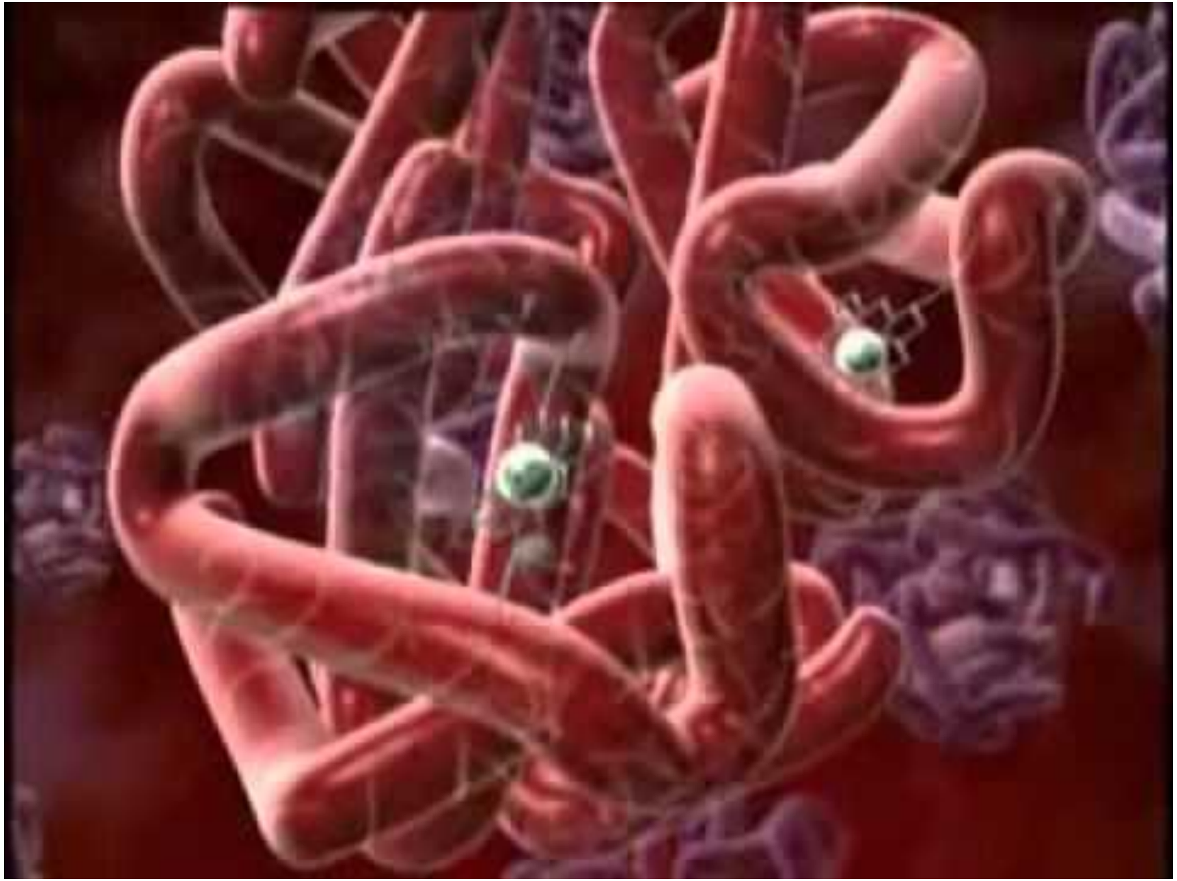
Decompose and build new



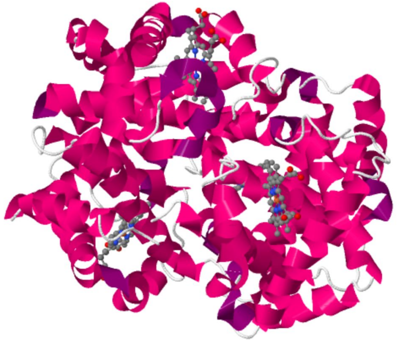
Proteins inside our body

Proteins are made up of amino acids MKTFLYFCLLFIVQTAFAADSIYVREQ

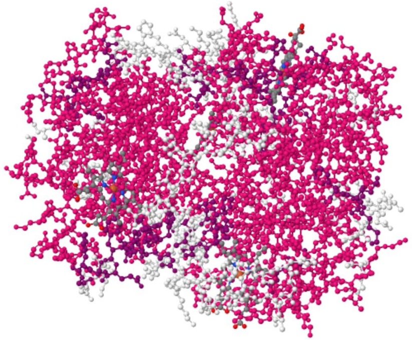
3D Structure Of Hemoglobin



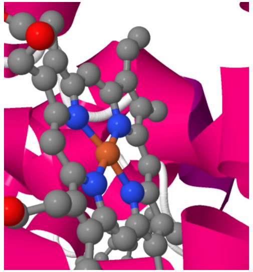
3D Structure Of Hemoglobin



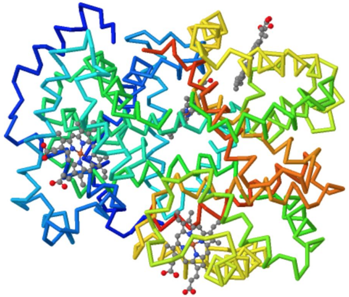
Cartoon representation



All atoms in a protein



Binding sites



Backbone representation

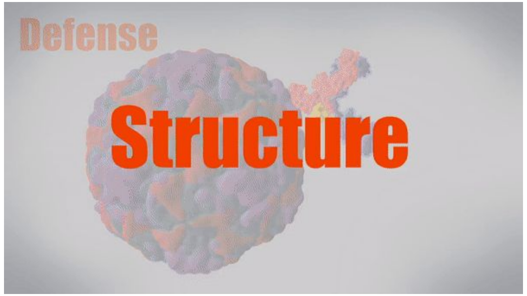
https://www.rcsb.org/pdb/explore/jmol.do?structureId=1LFZ&opt=3&bionumber=1

Functions of proteins (some examples)

Recognize and bind to foreign molecules – prevent viral DNA/RNA to enter the cell

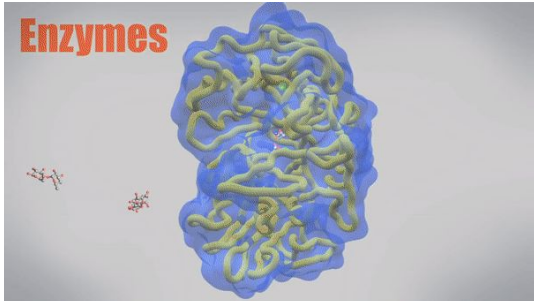
Alpha-amylase protein in saliva helps breaks down carbohydrates

Insulin regulates the blood sugar level



Collagen provides structural support – skin, cartilage, bones, etc.

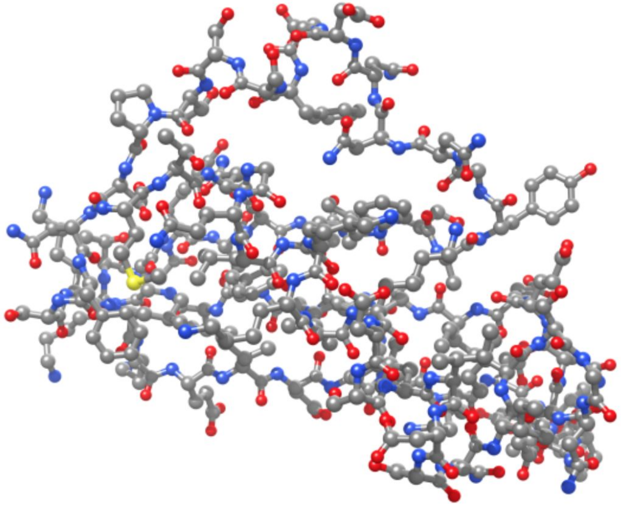
Calcium pumps allow synchronized muscles contraction



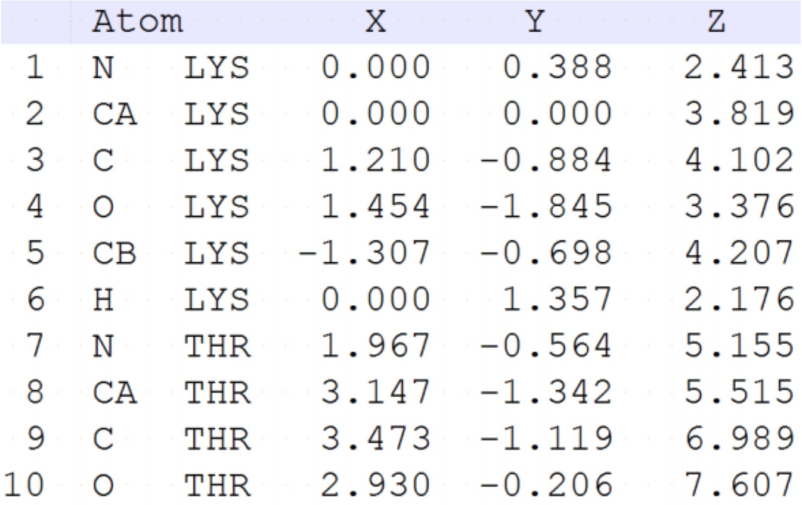
Ferritin forms a hollow shell to store iron from our food

All proteins are working, right now! And they are extremely busy!

Computer Representation



How biologists and chemists see a protein

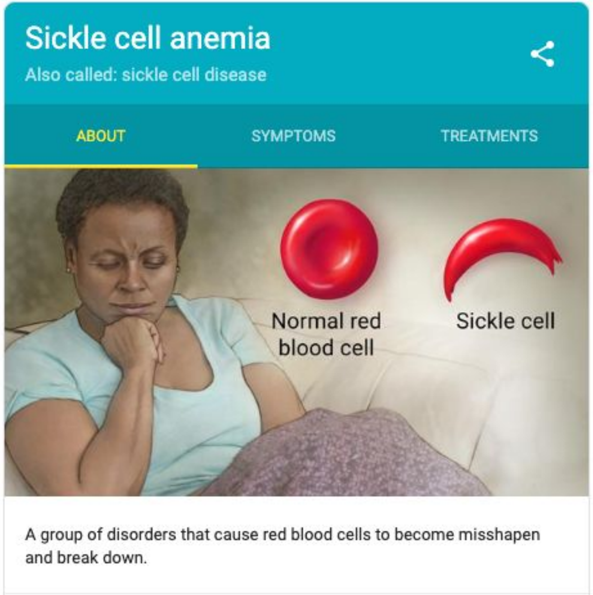


	Atom		X	Y	Z
1	N	LYS	0.000	0.388	2.413
2	CA	LYS	0.000	0.000	3.819
3	C	LYS	1.210	-0.884	4.102
4	O	LYS	1.454	-1.845	3.376
5	CB	LYS	-1.307	-0.698	4.207
6	H	LYS	0.000	1.357	2.176
7	N	THR	1.967	-0.564	5.155
8	CA	THR	3.147	-1.342	5.515
9	C	THR	3.473	-1.119	6.989
10	O	THR	2.930	-0.206	7.607

How computer scientists see a protein

Significance of Precise Structure

- Sickle cell anemia is caused by a mutation in a gene
  - The gene that tells our body to make the hemoglobin
- Sickle hemoglobin differs from normal hemoglobin by a single amino acid
  - *valine* replaces *glutamate* at position 6 on the surface of the beta chain



Sickle cell anemia

Also called: sickle cell disease

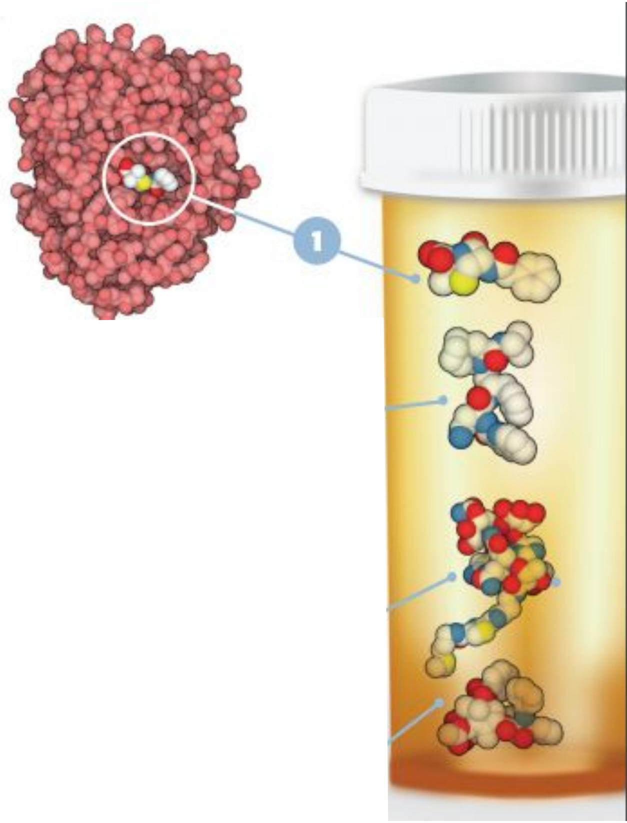
A group of disorders that cause red blood cells to become misshapen and break down.

The Need To Obtain Precise 3D Structures

- Antibiotics need to kill pathogenic organisms like bacteria without poisoning the patient

- Often, these drugs attack proteins that are only found in the targeted bacterium which are crucial for their survival or multiplication

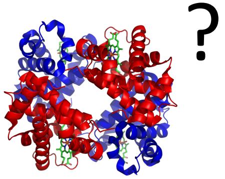
- For instance, penicillin attacks the enzyme that builds bacterial cell walls

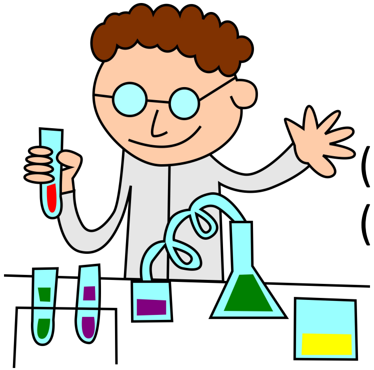


https://cdn.rcsb.org/pdb101/learn/resources/how-do-drugs-work-flyer.pdf

How Are 3D Structures Determined?

MKTFLYFCLLFIVQTAFAADSIYVREQ





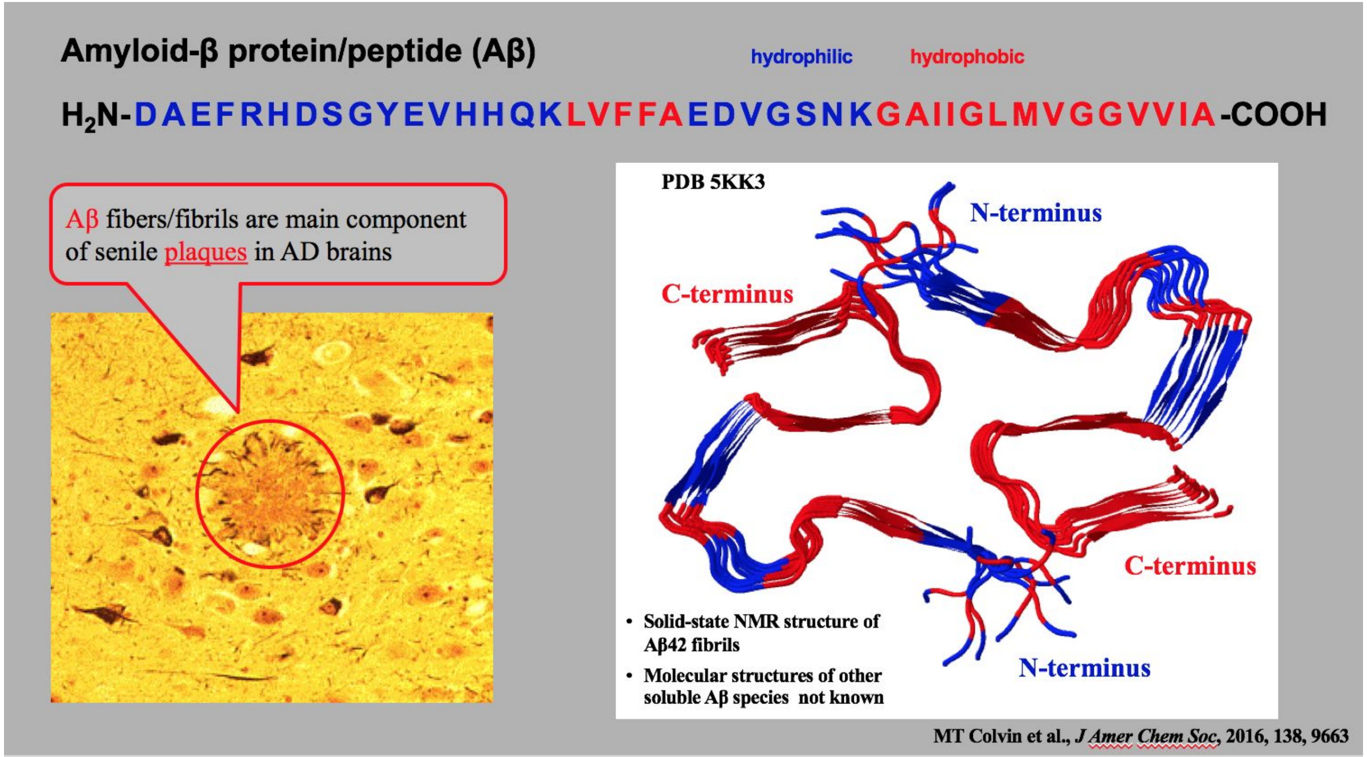
(1) NMR

(2) X-Ray Crystallography

- Accurate
- Takes months to years
- Expensive

Structure DB
(~ 200K proteins)
publicly available

Protein Oligomerization in Alzheimer's disease



Amyloid-β protein/peptide (Aβ)

hydrophilic hydrophobic

H_2N-DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVVIA-COOH

Aβ fibers/fibrils are main component of senile plaques in AD brains

- Solid-state NMR structure of Aβ42 fibrils
- Molecular structures of other soluble Aβ species not known

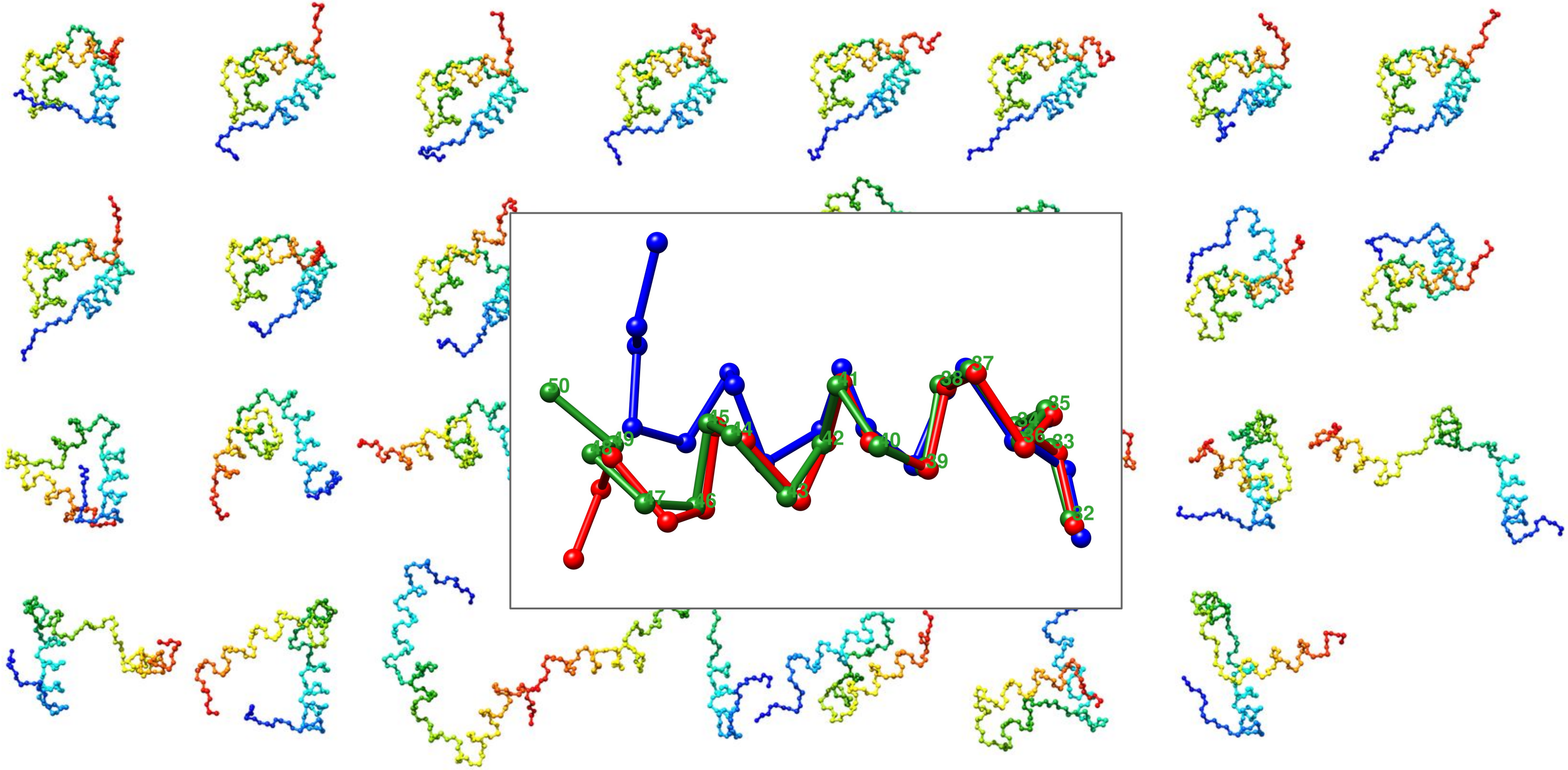
MT Colvin et al., *J Amer Chem Soc*, 2016, 138, 9663

"**These fibrils are of considerable medical importance** because they are **associated with more than 40 different diseases** including Alzheimer's disease."

Dr. Michael Nichols
Department of
Chemistry and Biochemistry
(UMSL)

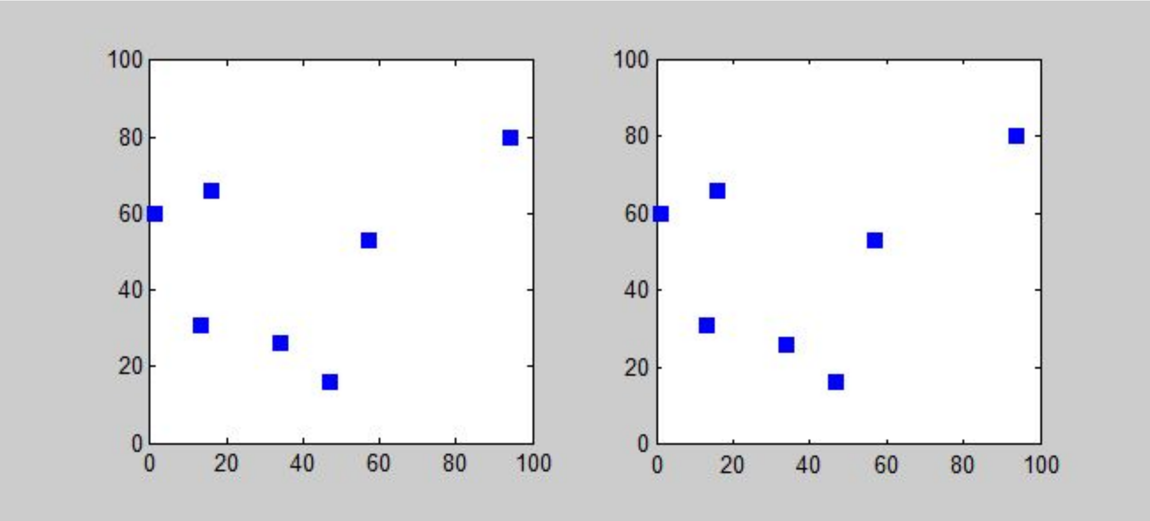
" .. my laboratory involves mechanistic studies of Aβ aggregation .. "

Protein Fold Space Search



A Similar Problem: Travelling Salesman

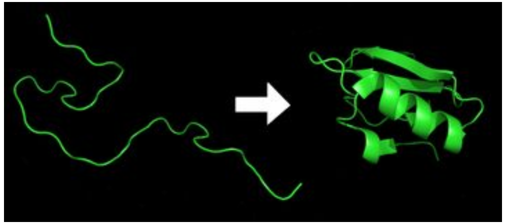
- Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and return to the origin city?

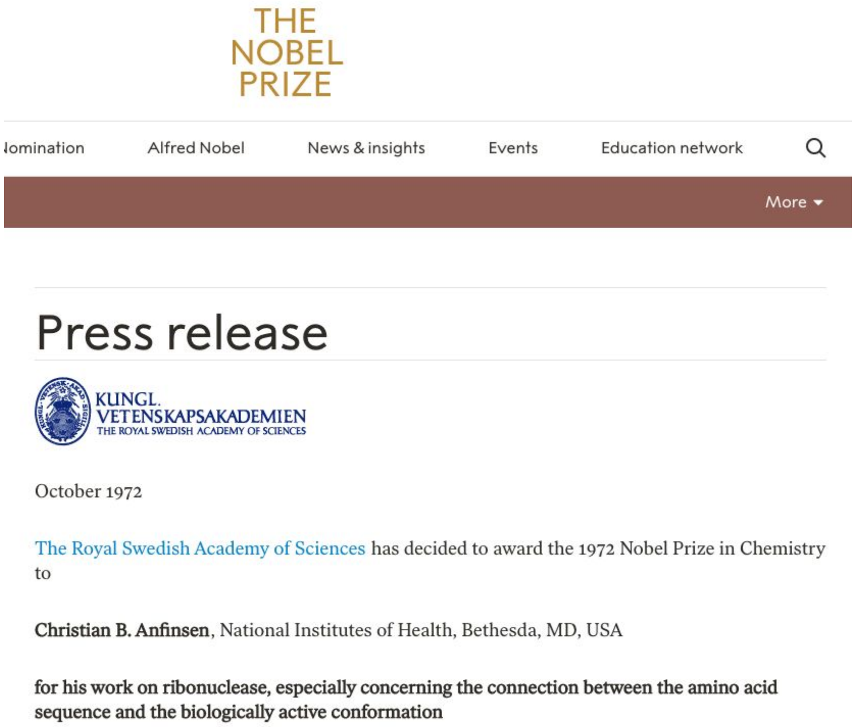


- Solution to a symmetric TSP with 7 cities using brute force search
    - Note: Number of permutations: (7-1)!/2 = 360

Anfinsen's Dogma (1973)

- Native structure is determined only by the protein's amino acid sequence





Press release

October 1972

The Royal Swedish Academy of Sciences has decided to award the 1972 Nobel Prize in Chemistry to

Christian B. Anfinsen, National Institutes of Health, Bethesda, MD, USA

for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation

Levinthal's Paradox

- In 1969, Cyrus Levinthal noted that, because of the very large number of degrees of freedom in an unfolded polypeptide chain, the molecule has an **astronomical number of possible conformations**

- If a protein were to attain its correctly folded configuration by sequentially sampling all the possible conformations, **it would require a time longer than the age of the universe to arrive at its correct native conformation**
    - This is true even if conformations are sampled at rapid (nanosecond or picosecond) rates

- The "paradox" is that most small proteins fold spontaneously on a millisecond or even microsecond time scale

Physics Based Approaches

- Disulfide Bonds
  - Disulfide bonds are formed between two sulfur (SH) atoms, which are found in the side-chain of the amino acid cysteine. When two cysteines are brought into close proximity..
- Electrostatic Bonds and Van der Waals Forces
  - Some amino acids have a charged side-chain, which is either negative or positive
  - Negatively charged side-chains are attracted to positively charged side-chains, while being repelled by another negatively charged side-chains
- Hydrogen Bonds
  - Hydrogen bonds form between two atoms and a hydrogen atom
  - Atoms, such as oxygen, can be covalently bound to hydrogen side-chain.
- Hydrophobic Interactions
  - Some amino acids have side-chains which repel water, or are hydrophobic

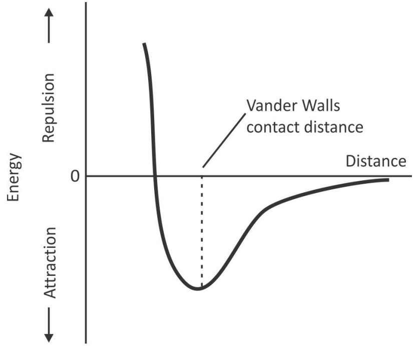


Fig. : Energy graph of Vander Wall's interaction.

Statistical / Knowledge-based Potentials

- A statistical potential or knowledge-based potential is an energy function derived from an analysis of known protein structures in the Protein Data Bank
    - For example, what is the likelihood that Glycine will interact with Glycine?

- An example of a highly successful method:
    - Rosetta

- Almost 40 years of research
    - And a lot of NSF/NIH money!

Experimentally or Computationally - We Obtain Restraints

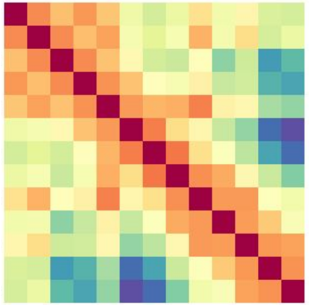
NMR / X-RAY

Algorithms / Deep Learning

G F G C N G P W D E D D M

Protein Sequence

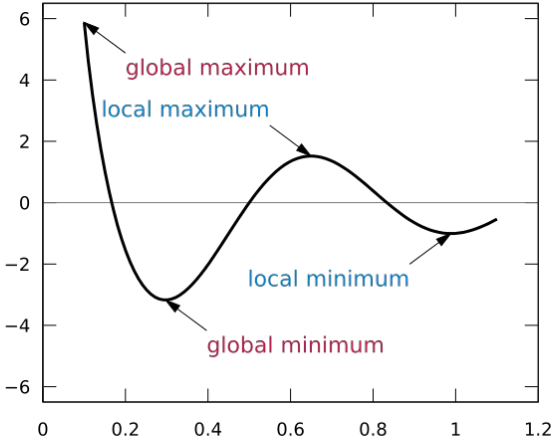
Predict which amino acids interact with which..



Distance Map

Energy Function

- What is energy function for TSP?
  - Sum of distances
- What is energy function for protein fold search?
  - ?



- Is the computation time needed for energy function evaluation important?

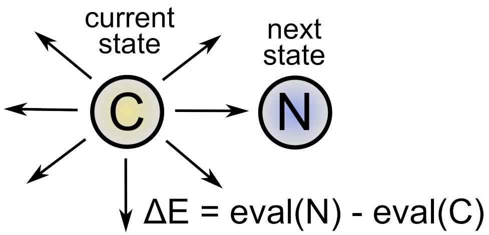
  Why?

- Can we compute energy of all configurations in the space?

  Can we list all the configurations at all?

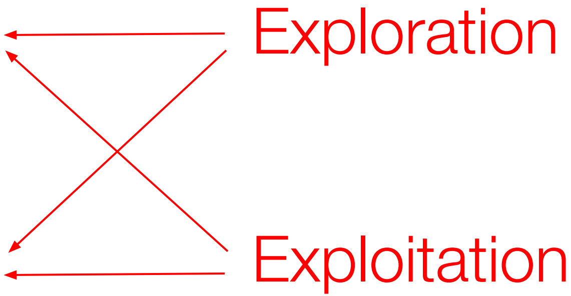
Hill Climbing vs Random Walk

1. Start with random values of the variables
2. Compute the energy function
3. Make small changes to the values, and see if we get a better value of the energy function
4. Repeat, until no better values of energy function can be obtained



```
while(next state is better than current){
    next $\leftarrow$ $moveGenerator(current)$
}
```

```
while (termination criterion is not reached){
    next $\leftarrow$ $randomMoveGenerator(current)$
}
```


Exploration

Exploitation

The Simulated Annealing Algorithm

Annealing

- physical process of controlled cooling

Why?

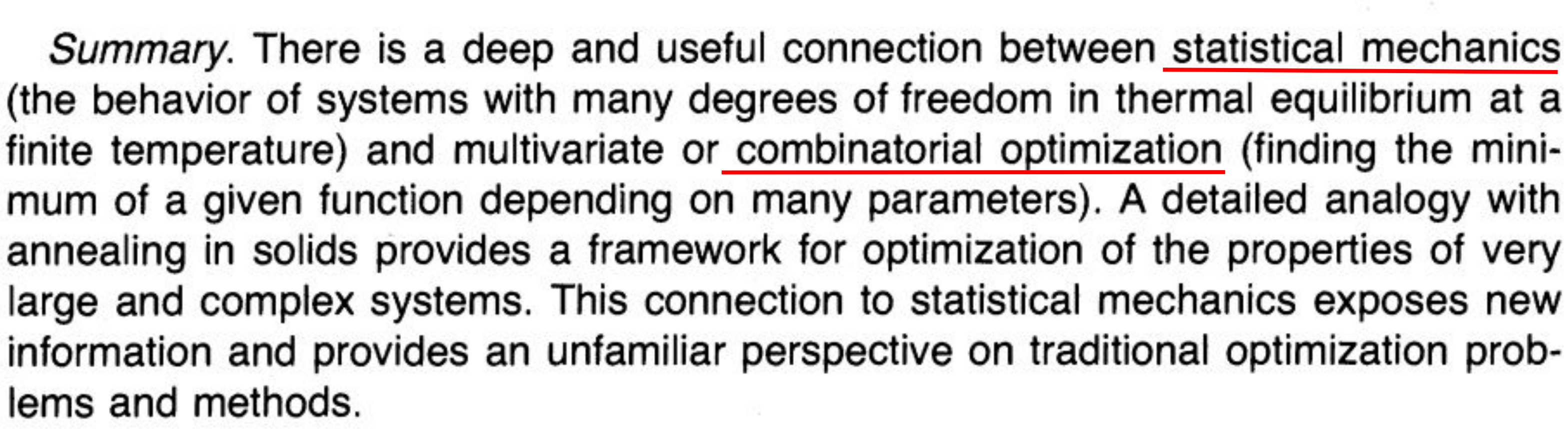
- Produce materials of good properties, like strength
- Involves create liquid version and then solidifying
    - Example: casting

Intuition (we want minimum energy)

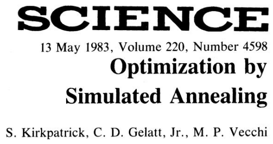
- Desirable to arrange the atoms in a systematic fashion, which in other words corresponds to low energy

The Simulated Annealing Algorithm

The Original paper:



Summary. There is a deep and useful connection between statistical mechanics (the behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature) and multivariate or combinatorial optimization (finding the minimum of a given function depending on many parameters). A detailed analogy with annealing in solids provides a framework for optimization of the properties of very large and complex systems. This connection to statistical mechanics exposes new information and provides an unfamiliar perspective on traditional optimization problems and methods.



SCIENCE

13 May 1983, Volume 220, Number 4598

Optimization by Simulated Annealing

S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi

The Simulated Annealing Algorithm

- Statistical Mechanics
  - The behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature.

- Combinatorial Optimization
  - Finding the minimum of a given function depending on many variables.

- The Analogy
  - If a liquid material cools and anneals too quickly, then the material will solidify into a sub-optimal configuration
  - If the liquid material cools slowly, the crystals within the material will solidify optimally into a state of minimum energy (i.e. ground state)
  - This ground state corresponds to the minimum of the cost function in an optimization problem

Simulated Annealing: Intuition

1TapBubbles Water Ring Toss LT

1Tapps - The Shortcuts Company - One Tap Apps - December 23, 2014

Board

This app is compatible with all of your devices.

★★★☆☆ (153)

+74 Recommend this on Google

Labyrinth Lite

Illusion Labs - September 13, 2012

Arcade

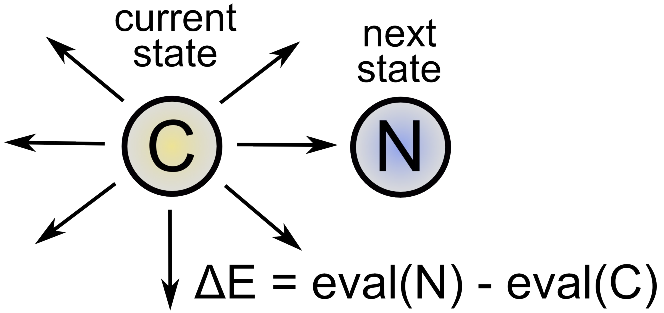
This app is compatible with all of your devices.

★★★★☆ (82,703)

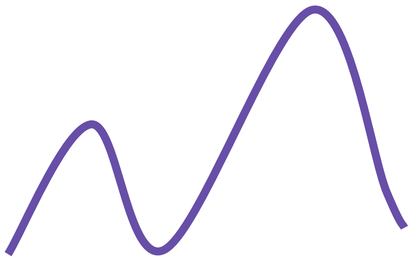
In simulations, we can roll back!

We Are Confident With Good Moves = Probability

- Make move with some probability
    - such that, if the move is a good one, the probability is high



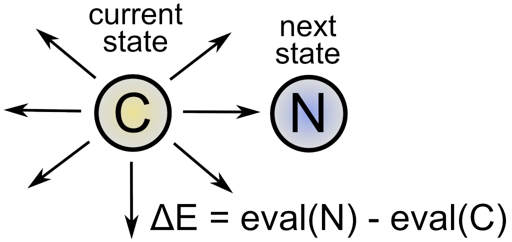
ΔE = eval(N) - eval(C)



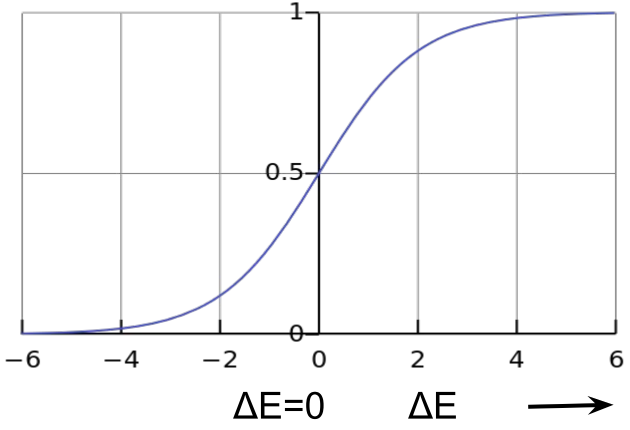
- We want ΔE to influence the probability
    - allow good moves (high ΔE) with high probability, and allow bad moves with low probability

Control How ΔE Influences Probability

- What function is appropriate?
    - Range should be 0 and 1 (probability)
    - Domain should be infinite



$$P(C,N) = \frac{1}{1+e^{-\Delta E}}$$



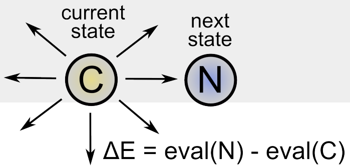
Beginning vs End (Level of Carefulness) = Temperature

T can control how ΔE influences probability

$$P(C,N) = \frac{1}{1+e^{-\Delta E}} \longrightarrow P(C,N) = \frac{1}{1+e^{\frac{-\Delta E}{T}}}$$

How do ΔE and T influence the probability of accepting a move?

Effect of ΔE and T on Probability



$$P(C,N) = \frac{1}{1+e^{\frac{-\Delta E}{T}}}$$

effect of $\Delta\mathrm{E}$:

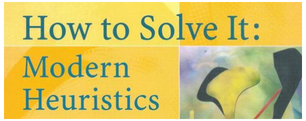
$T = 10, eval(C) = 107$

$\mathbf{eval(N)}$	$\mathbf{-\Delta E}$	$\mathbf{e^{-\frac{\Delta E}{T}}}$	$\mathbf{P}$
80	27	14.88	0.06
100	7	2.01	0.33
107	0	1.0	0.50
120	-13	0.27	0.78
150	-43	0.01	0.99

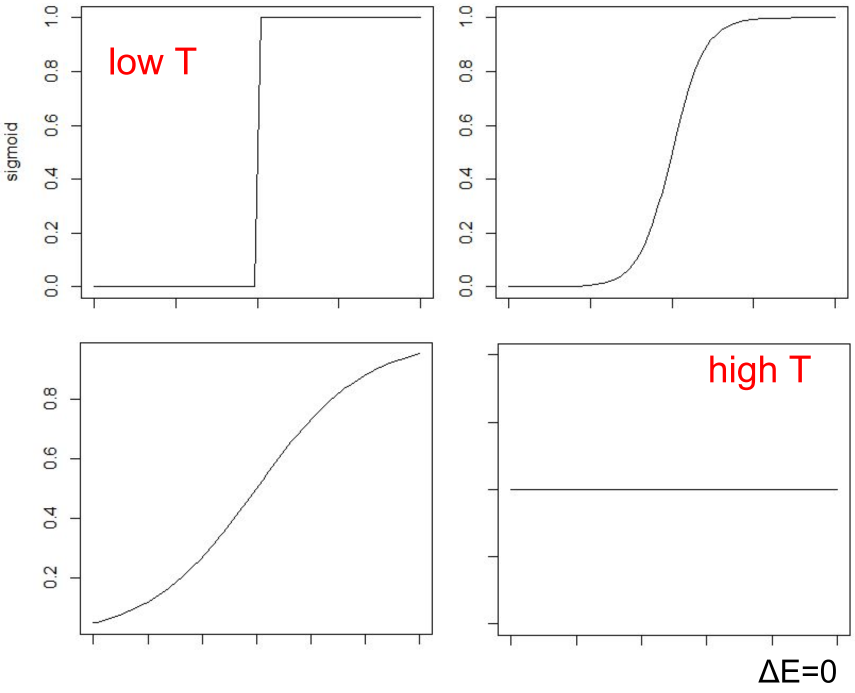
worse ... better

effect of T :

$\mathbf{T}$	$\mathbf{e^{-\frac{13}{T}}}$	$\mathbf{P}$	
1	0.000002	1.0	HC
5	0.074	0.93	
10	0.27	0.78	
20	0.52	0.66	
50	0.77	0.56	
10^{10}	0.9999	0.5	RW


How to Solve It: Modern Heuristics

Effect of Temperature (contd.)

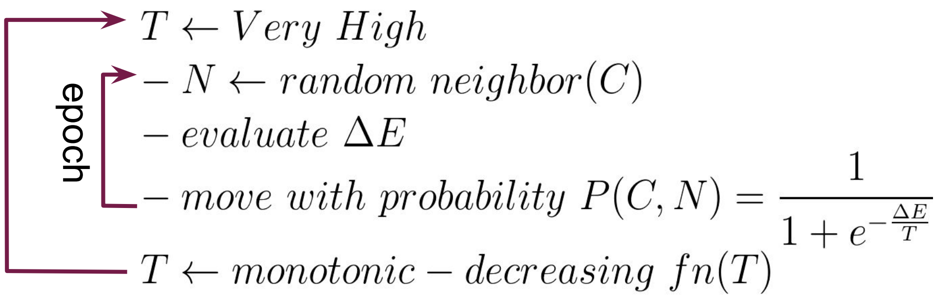


sigmoid function

effect of T :

$\mathbf{T}$	$\mathbf{e^{-\frac{13}{T}}}$	$\mathbf{P}$
1	0.000002	1.0
5	0.074	0.93
10	0.27	0.78
20	0.52	0.66
50	0.77	0.56
10^{10}	0.9999	0.5

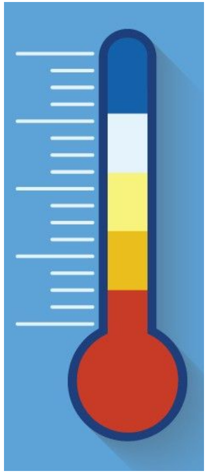
The Simulated Annealing Algorithm



```
$T \leftarrow Very\ High$
    $- N \leftarrow random\ neighbor(C)$
    $- evaluate\ \Delta E$
    $- move\ with\ probability\ P(C,N) = \frac{1}{1+e^{-\frac{\Delta E}{T}}}$
$T \leftarrow monotonic - decreasing\ fn(T)$
```

How To Choose T?

- We follow the physical world approach
  - cool the system gradually and hope that the system will settle to an optimal state
- We initialize T to some high value
  - Gradually decrease T
  - Some monotonically decreasing function (cooling rate)
- How to select initial and final temperature?
  - Empirically



Key Ingredients of the Simulated Annealing Algorithm

1. A concise description of a **configuration** (architecture, design, topology) of the system (Design Vector)
2. A **random generator of rearrangements** of the elements in a configuration (Neighborhoods). This generator encapsulates rules so as to generate only valid configurations
3. **Perturbation function**. A quantitative objective function containing the trade-offs that have to be made (Simulation Model and Output Metric(s)). Surrogate for system energy
4. An **annealing schedule of the temperatures** and/or the length of times for which the system is to be evolved.

Typical Problems (during implementation)

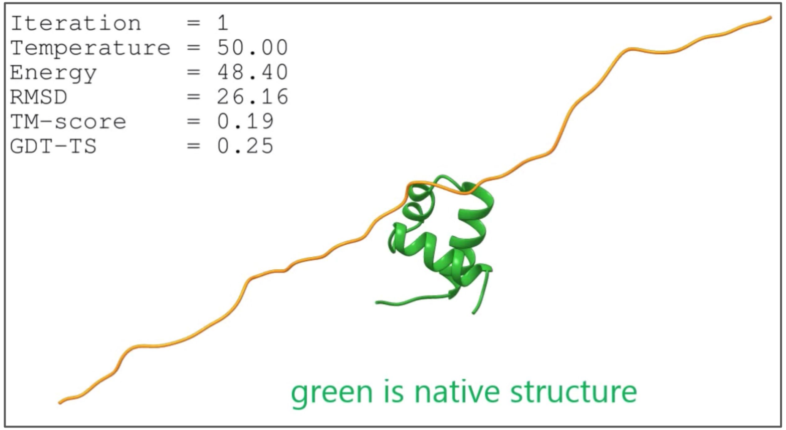
- Initial temperature is too high
- Temperature goes down too quickly
- Process completes and the solution is not optimal

Limitations

- For SA to work satisfactorily, the cost function should not contain narrow and steep valleys
  - http://www.drdobbs.com/architecture-and-design/simulated-annealing-a-heuristic-optimiza/184404780
- For problems where the energy landscape is smooth, or there are few local minima, SA is overkill
  - simpler, faster methods (e.g., gradient descent) will work better
- Heuristic methods, which are problem-specific or take advantage of extra information about the system, will often be better than general methods, although SA is often comparable to heuristics
- The method cannot tell whether it has found an optimal solution
  - Some other complementary method (e.g. branch and bound) is required to do this https://cs.adelaide.edu.au/~paulc/teaching/montecarlo/node140.html

Reconstruction Of Protein Molecule ‘1GUU’

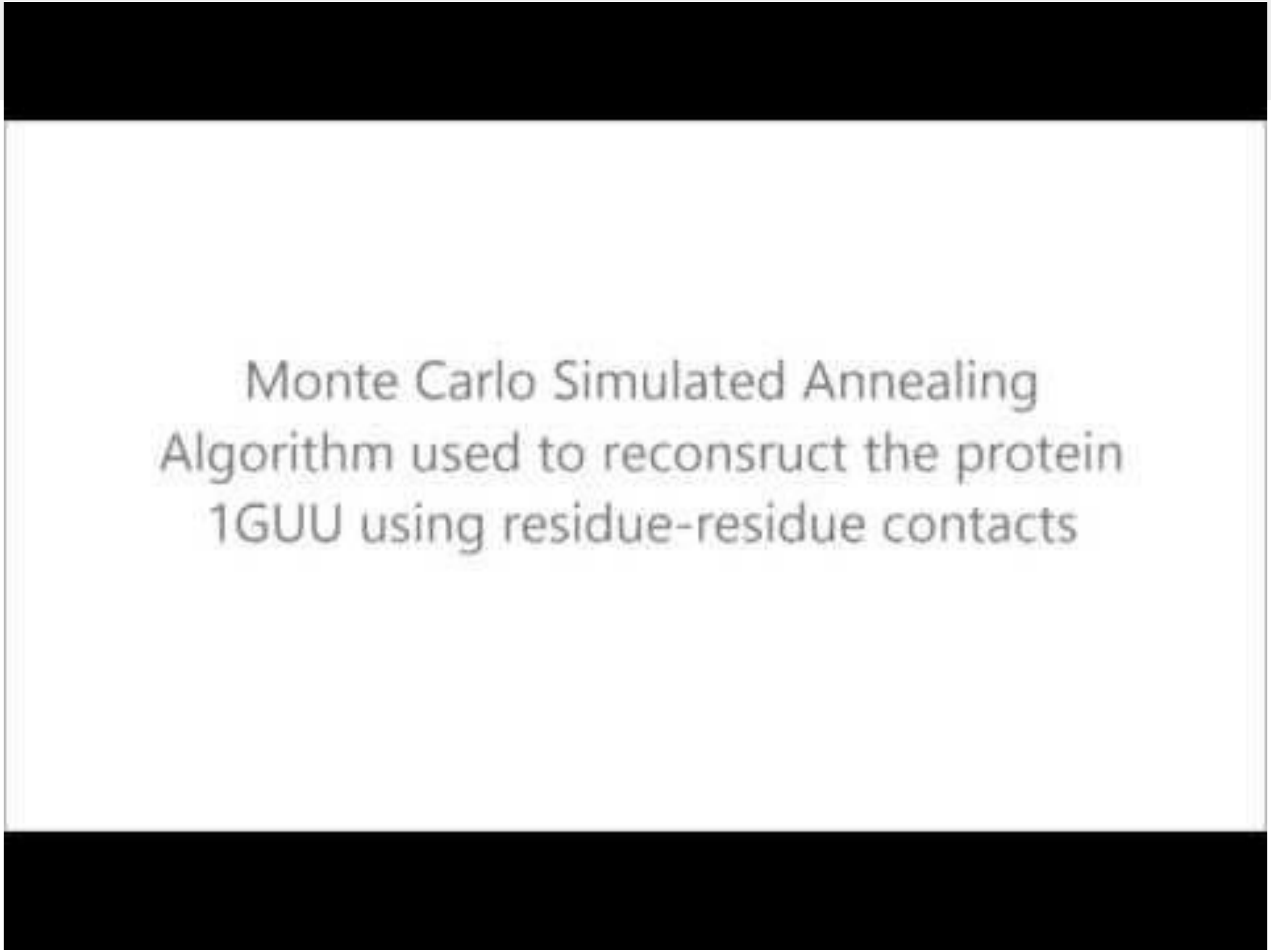
- Initial Temperature = 50 (length of the protein)
- Final Temperature = 0.01
- Number of SA iterations = 3200
- Number of true Cβ contacts = 45
- Contact energy = Root Mean Square Deviations
- T_i+1 = T_i - 10
- N_{epoch} = 100



```
Iteration   = 1
Temperature = 50.00
Energy      = 48.40
RMSD        = 26.16
TM-score    = 0.19
GDT-TS      = 0.25
```

green is native structure

A movie clip demonstrating an application of monte carlo simulated annealing to reconstruct a small alpha helical protein 1GUU.
https://www.youtube.com/watch?v=2p5x7XROxlo



Monte Carlo Simulated Annealing Algorithm used to reconsruct the protein 1GUU using residue-residue contacts

THANK YOU