# Protein Structure Prediction

## CPSC 445

Chris Thachuk

Guest Lecture, March 27th 2007

# Learning Objectives

- basic molecular biology
- history of protein structure determination
- overview of popular methods for structure determination
- appreciation for simplified protein models

# What is a protein?

## Definition

A protein, is a chain of amino acids coded for by a gene using the genetic code. Proteins serve the basis for structure, function and communication within and between living cells. They provide the mechanisms for cell growth and cell reproduction.

# What is a protein?

## Definition

A protein, is a chain of amino acids coded for by a gene using the genetic code. Proteins serve the basis for structure, function and communication within and between living cells. They provide the mechanisms for cell growth and cell reproduction.

## Classes of Proteins

Enzymes, Receptors, Replicases and Polymerases, Hormones, Motor proteins, Structural proteins, . . .

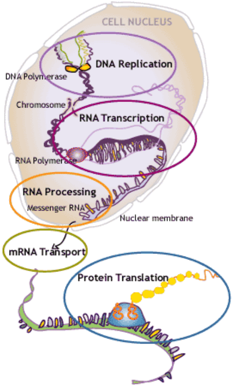
# What is a protein?

## Definition

A protein, is a chain of amino acids coded for by a gene using the genetic code. Proteins serve the basis for structure, function and communication within and between living cells. They provide the mechanisms for cell growth and cell reproduction.

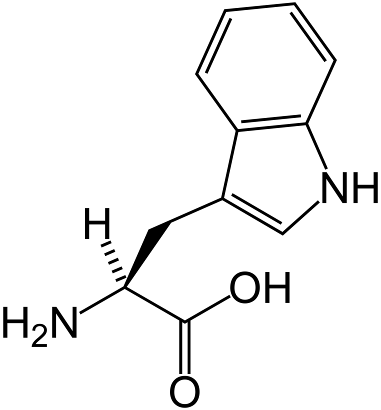
## Classes of Proteins

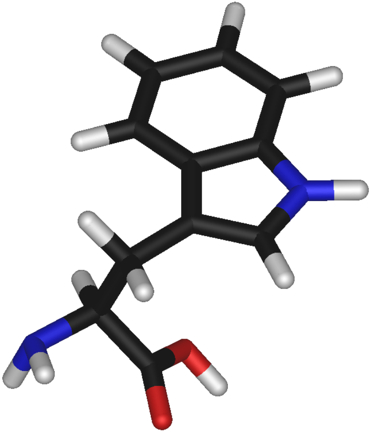
Enzymes, Receptors, Replicases and Polymerases, Hormones, Motor proteins, Structural proteins, . . .



Nobel Prize Website - [http://nobelprize.org]

# What is an amino acid?





Wikimedia Commons - [http://wikipedia.org]

# Some Protein Facts

- Protein: from the greek word prota meaning of primary importance

Nobel Prize Website - [http://nobelprize.org]

# Some Protein Facts

- Protein: from the greek word prota meaning of primary importance
- First described in 1838 by Jons Jakob Berzelius

Nobel Prize Website - [http://nobelprize.org]

# Some Protein Facts

- Protein: from the greek word prota meaning of primary importance
- First described in 1838 by Jons Jakob Berzelius
- First protein was crystalized in 1926 by James Sumner

Nobel Prize Website - [http://nobelprize.org]

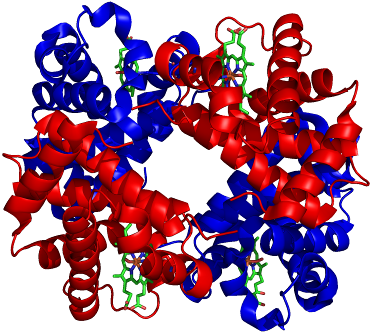
# Some Protein Facts

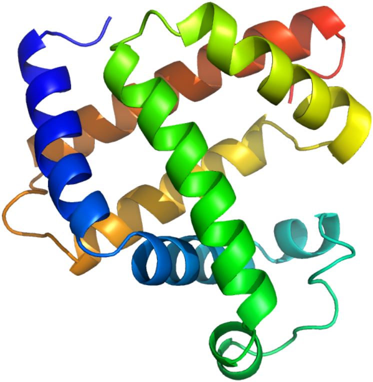
- Protein: from the greek word prota meaning of primary importance
- First described in 1838 by Jons Jakob Berzelius
- First protein was crystalized in 1926 by James Sumner
- First protein sequenced was Insulin by Frederick Sanger in 1955

Nobel Prize Website - [http://nobelprize.org]

# Some Protein Facts (Cont'd)

- First structures solved were Myoglobin and Haemoglobin by Sir John Cowdery Kendrew and Max Perutz in 1962



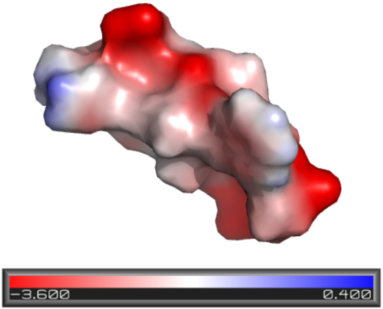


Protein Data Bank - [http://pdg.org]

# Motivation: Why is protein structure important?

## Accepted Dogma

`...form gives rise to function...`



Produced with MacPyMol - [http://delsci.com/macpymol]

Landliving - [http://www.landliving.com]

# Far reaching implications

## If:

- It is easy to deduce function from form, and
- we could determine the form of entire proteomes

# Far reaching implications

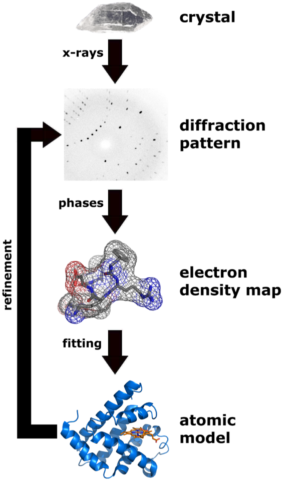
## If:

- It is easy to deduce function from form, and
- we could determine the form of entire proteomes

## Then:

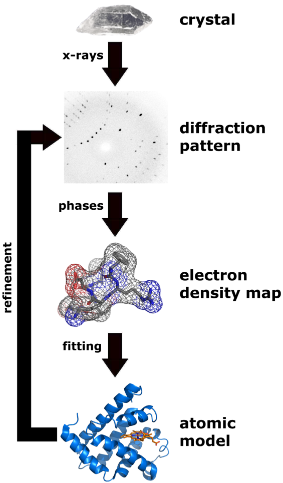
- We would gain tremendous insight into countless diseases
- Learn how to treat these condition
- Learn more about molecular biology in general
- etc . . .

# Determing Structure: X-Ray Crystallography



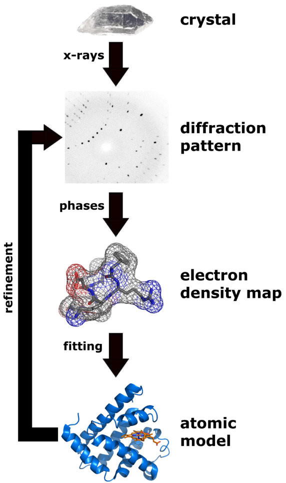
- A crystal of the protein being studied must first be created

# Determing Structure: X-Ray Crystallography



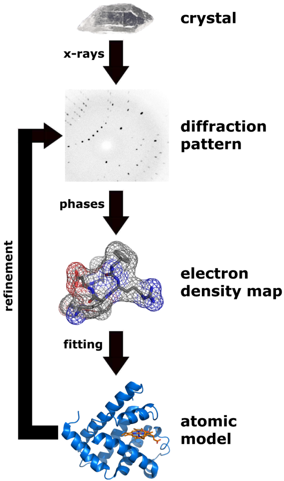
- A crystal of the protein being studied must first be created
- Based on the concept of diffraction
  - Why X-rays?
  - Why diffraction?

# Determing Structure: X-Ray Crystallography



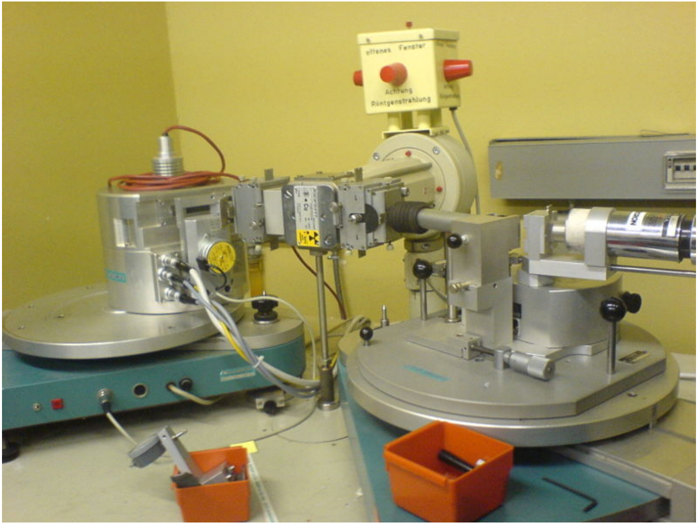
- A crystal of the protein being studied must first be created
- Based on the concept of diffraction
  - Why X-rays?
  - Why diffraction?
- A model can then be constructed to explain the diffraction pattern

# Determing Structure: X-Ray Crystallography



- A crystal of the protein being studied must first be created
- Based on the concept of diffraction
  - Why X-rays?
  - Why diffraction?
- A model can then be constructed to explain the diffraction pattern
- Refinement of the model can ensue

# Determing Structure: X-Ray Crystallography (2)



# Determing Structure: NMR Spectroscopy

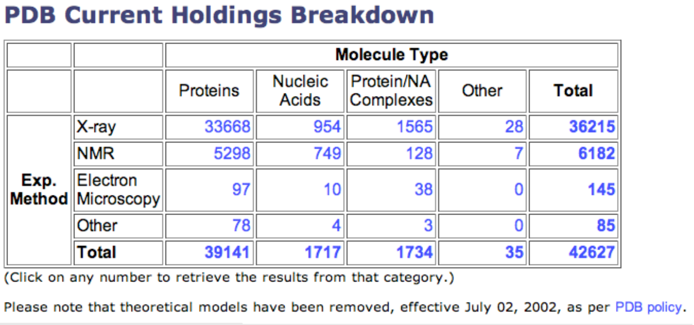
## NMR

Nuclear magnetic resonance attempts to capture the structure and dynamics of proteins by using radio frequency pulses and detecting delays of transfer between neighbouring nuclei.

Pacific Northwest National Laboratories NMR facility - [public domain]

# Determining Structure: Current Structure Knowledge

The Protein Data Bank (http://pdb.org) is currently the standard repository for 3-dimensional structure models.



## PDB Current Holdings Breakdown

| | | Molecule Type | | | | |
|---|---|---|---|---|---|---|
| | | Proteins | Nucleic Acids | Protein/NA Complexes | Other | **Total** |
| **Exp. Method** | X-ray | 33668 | 954 | 1565 | 28 | **36215** |
| | NMR | 5298 | 749 | 128 | 7 | **6182** |
| | Electron Microscopy | 97 | 10 | 38 | 0 | **145** |
| | Other | 78 | 4 | 3 | 0 | **85** |
| | **Total** | **39141** | **1717** | **1734** | **35** | **42627** |

(Click on any number to retrieve the results from that category.)

Please note that theoretical models have been removed, effective July 02, 2002, as per PDB policy.

*2008 Update: 45,906 (proteins)*

# Okay, Let’s Determine All Structures!

Wikimedia Commons - [http://commons.wikimedia.org]

# Well . . .

## As it turns out:

- Lab techniques are often prohibitively expensive
- Some protein structures are impossible to determine by current techniques
- Can be very time consuming process

# Some Hope

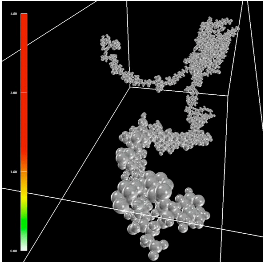
## Definition

Anfinsen's principle dictates that the information necessary to determine the unambiguous three dimensional structure of a protein is contained within the polypeptide chain.

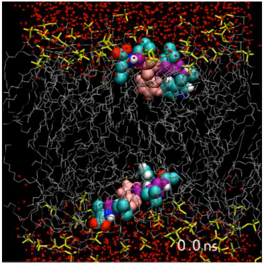
- Suggested by Christian Anfinsen in the 1960's
- Widely accepted by the community

# ab initio (from first principles)

Using only the primary sequence information, deduce the 3D structure of the protein. Often determined by a Monte Carlo search method or molecular dynamics simulation



The AMBER empirical energy of protein t209 - [Ernest Orlando Lawrence Berkeley National Laboratory - Visualization Group]



Trajectory of the Phe-peptide in the DOPC bilayer - [http://moose.bio.ucalgary.ca]

# ab initio (from first principles) cont'd

## Some Resources

- **Rosetta Commons** (http://www.rosettacommons.org) - resources from many groups including Dr. David Baker at University of Washington. Specifically RosettaAbInitio.

- **Robetta Server** (http://robetta.org) - online public server front end to rosetta software.

# Homology Modeling

Start with the known structure of a homologous protein and refine the model. Often, the sequence is threaded onto the backbone of the known structure (ie. protein threading). This is usually the most accurate theoretical prediction method.

## Some Resources

- **Swiss Model** (http://expasy.org/swissmod/SWISS-MODEL.html) - performs homology search to determine suitable starting structure if possible.

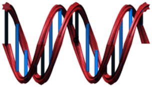
# Simplified Models (out of necessity)

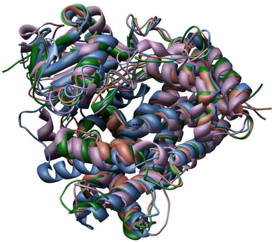
## Levinthal's Paradox

The ensemble of conformations for a protein is astronomical. However, a protein always manages to fold consistently, into a unique structure in the order of milli-seconds to seconds.

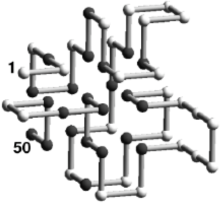
Simplified models have been developed to study the overall folding process and are used to provide a starting point for more accurate methods.

# In Vivo vs. In Silico





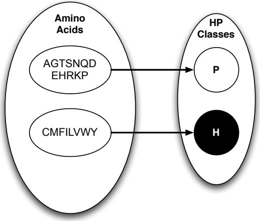
**actcg**



# The HP Model

## Definition

The hydrophobic polar (HP) model maps each of the 20 amino acids to one of two classes: hydrophobic or polar.



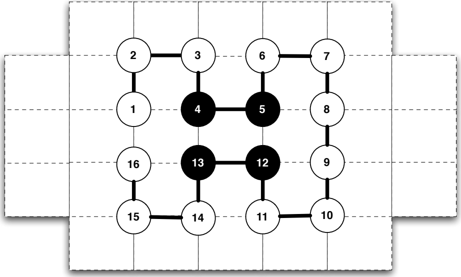
Dill KA: **Theory for the folding and stability of globular proteins**. Biochemistry 1985, 24(6):1501-1509. 37.

Lau KF, Dill KA: **A lattice statistical mechanics model of the conformational and sequence spaces of proteins**. Macromolecules 1989, 22(10):3986-3997.

# Lattices and Energy Minimization

## Assumption (Model Simplification)

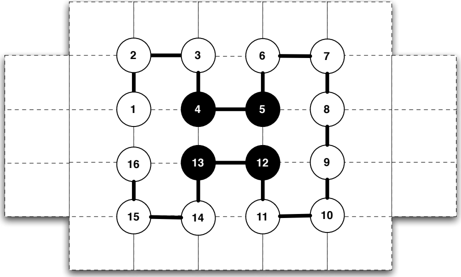
A native conformation contains the maximum possible number of hydrophobic-hydrophobic (H-H) contacts between non-neighbouring amino acids.



# Lattices and Energy Minimization

## Assumption (Model Simplification)

A native conformation contains the maximum possible number of hydrophobic-hydrophobic (H-H) contacts between non-neighbouring amino acids.



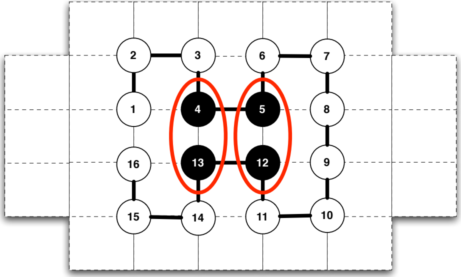
## Question

How many non-neighbour H-H contacts in this conformation?

# Lattices and Energy Minimization

## Assumption (Model Simplification)

A native conformation contains the maximum possible number of hydrophobic-hydrophobic (H-H) contacts between non-neighbouring amino acids.



## Question

How many non-neighbour H-H contacts in this conformation?

## Answer

2

# Minimum Free Energy

## Energy Function

$$E(c_i) = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} N_{jk}, \text{ where} \quad (1)$$

$$N_{jk} = \begin{cases} -1 & \text{if } j \text{ and } k \text{ are both H residues} \\ & \text{and } \textit{topological neighbours}; \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

# Minimum Free Energy

## Energy Function

$$E(c_i) = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} N_{jk}, \text{ where} \tag{1}$$

$$N_{jk} = \begin{cases} -1 & \text{if } j \text{ and } k \text{ are both H residues} \\ & \text{and } \textit{topological neighbours}; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

## Minimum Free Energy

$$E(c^*) = min\{E(c_i)|c_i \in C_s\} \tag{3}$$

# Complexity of Protein Folding

## This is a hard problem

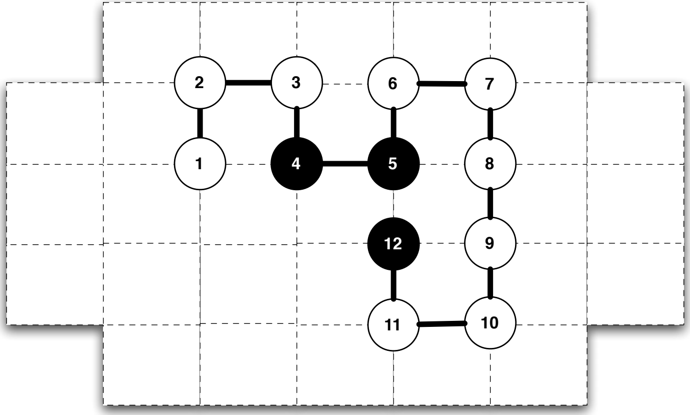
Folding proteins has been shown to be $\mathcal{NP}$-hard, even for our simplified models restricted to lattices.

Berger B, Leighton T: **Protein folding in the hydrophobic-hydrophilic (HP) is NP-complete**. In Proceedings of the second annual international conference on Computational molecular biology 1998:30-39.

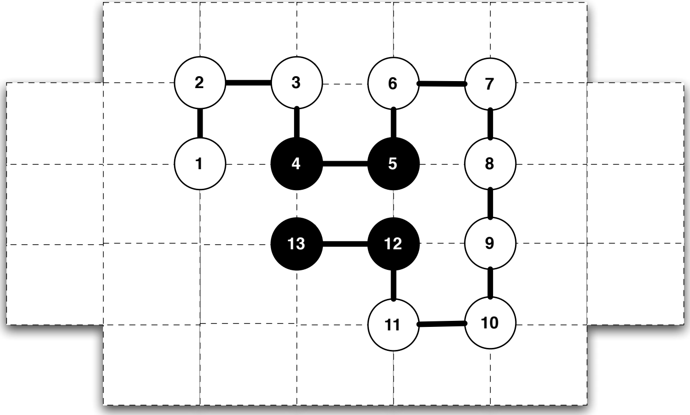
Crescenzi P, Goldman D, Papadimitriou C, Piccolboni A, Yannakakis M: **On the complexity of protein folding**. In Proceedings of the second annual international conference on Computational molecular biology 1998:61-62.

Hart W, Istrail S: **Robust proofs of NP-hardness for protein folding**: general lattices and energy potentials. Journal of Computational Biology 1997, 4.

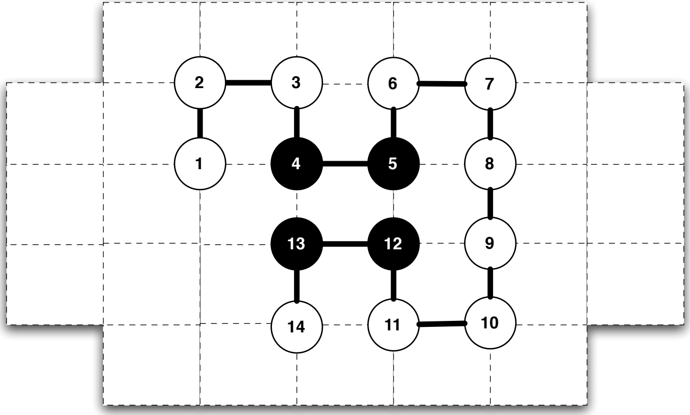
# Chain Growth Algorithms



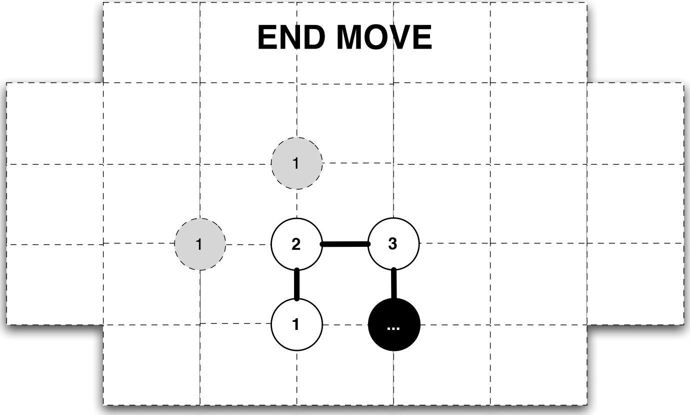
# Chain Growth Algorithms



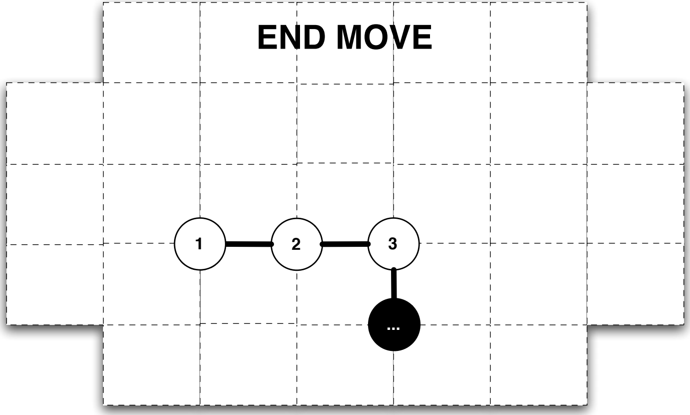
# Chain Growth Algorithms



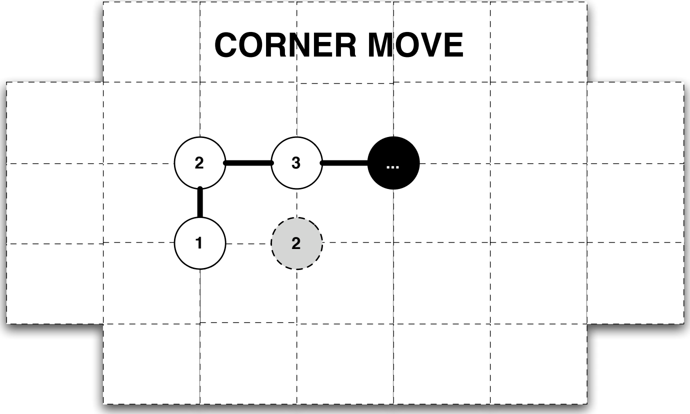
# Local Search Algorithms (Local Moves)



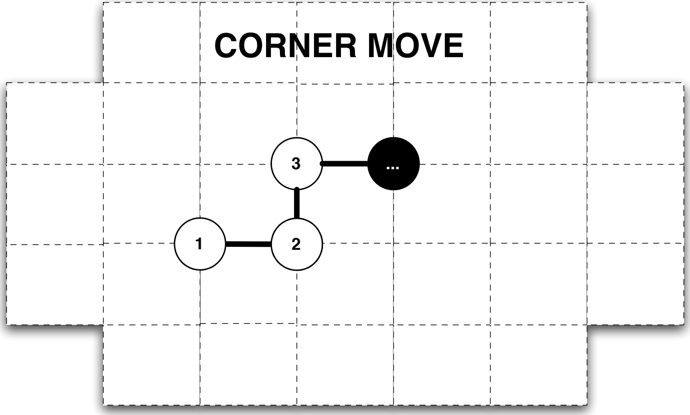
# Local Search Algorithms (Local Moves)



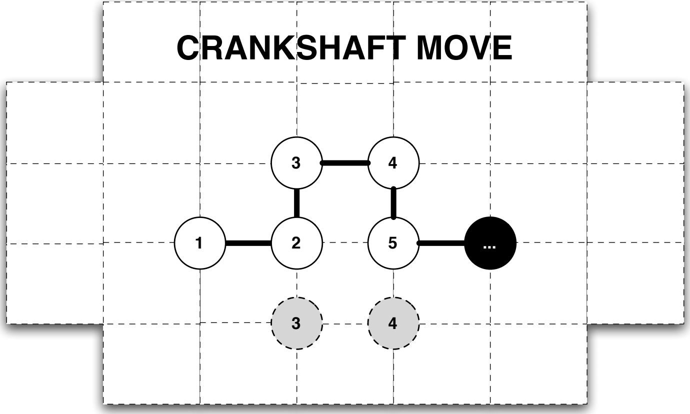
# Local Search Algorithms (Local Moves)



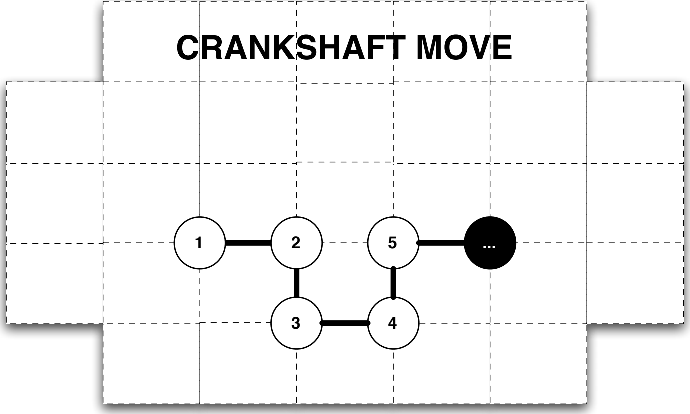
# Local Search Algorithms (Local Moves)



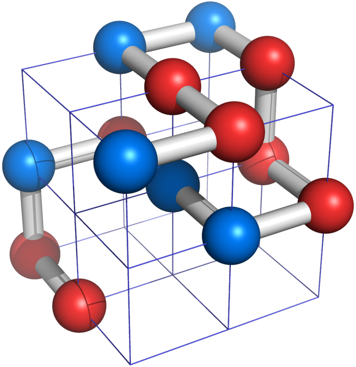
# Local Search Algorithms (Local Moves)

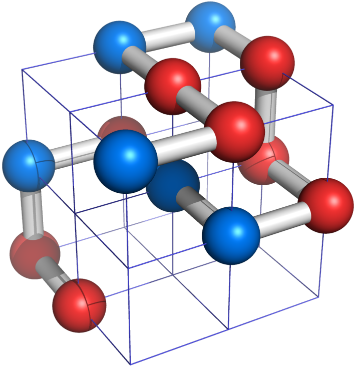


# Local Search Algorithms (Local Moves)



# Chain Growth and Local Moves in 3D





# Existing Algorithms

## PERM

- Chain growth algorithm
- Generates candidate solutions, prunes and enriches
- ~~Currently~~ Was state of the art algorithm for 2D/3D HP Model
- Inherent difficulty with folds involving interacting termini

Grassberger P: **Pruned-enriched Rosenbluth method: Simulations of $\theta$ polymers of chain length up to 1 000 000**. Phys. Rev. E 1997, 56(3):3682-3693.

# Existing Algorithms (2)

## ACOHPPFP-3

- Chain growth & local search phases
- Ant Colony Optimization search algorithm (stymergy)
- Larger range of contact order values than PERM
- No difficulty with folds involving interacting termini

Shmygelska A, Hoos H: **An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem**. BMC Bioinformatics 2005, 6:30.

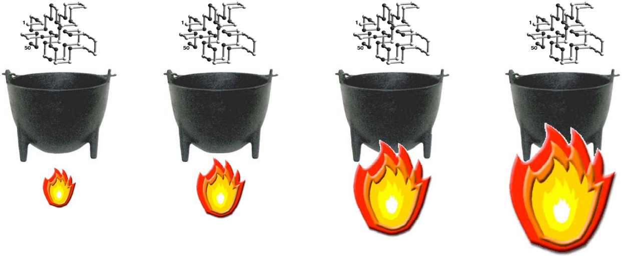
# Existing Algorithms (3)

## Replica exchange Monte Carlo

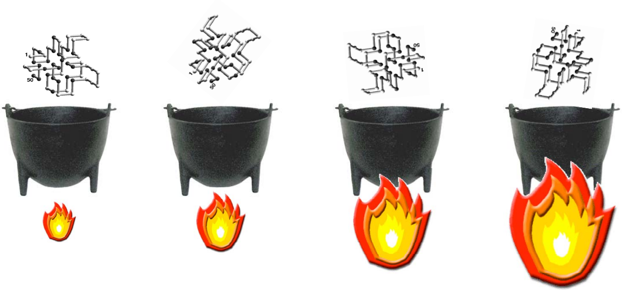
- Discovered independently by three groups
- Extended ensemble algorithm
- Previous success in rough landscapes
- Applied to off-lattice protein folding

Thachuk C, Shmygelska A, Hoos H: **Replica exchange Monte Carlo for protein folding in the HP model**. BMC Bioinformatics 2007, 8:342.

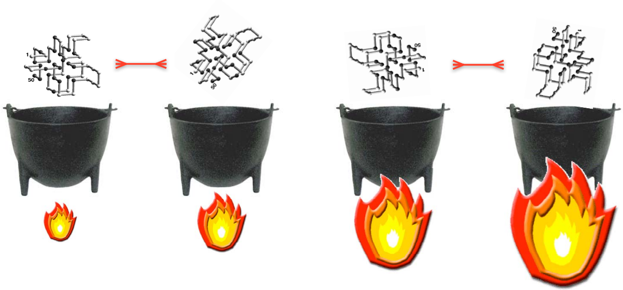
# Replica Exchange - The general idea



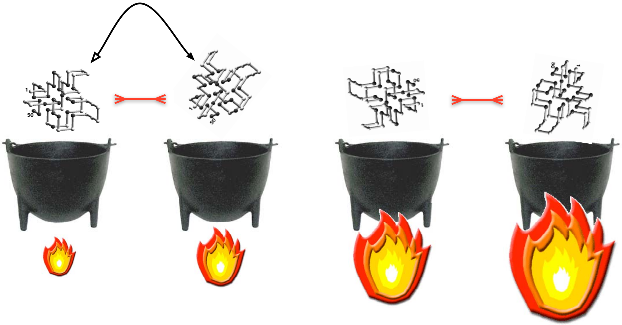
# Replica Exchange - The general idea



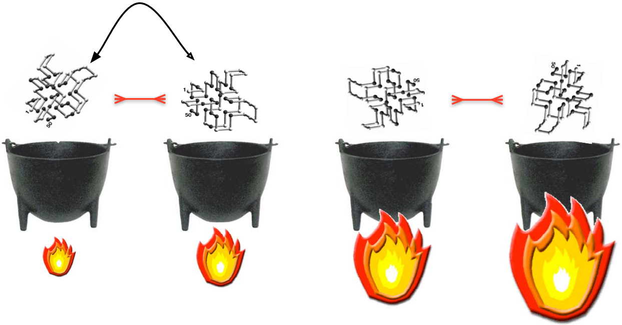
# Replica Exchange - The general idea



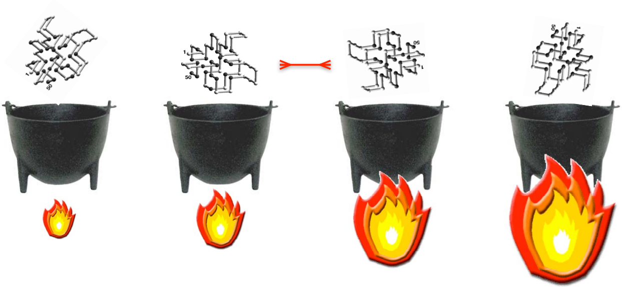
# Replica Exchange - The general idea



# Replica Exchange - The general idea



# Replica Exchange - The general idea

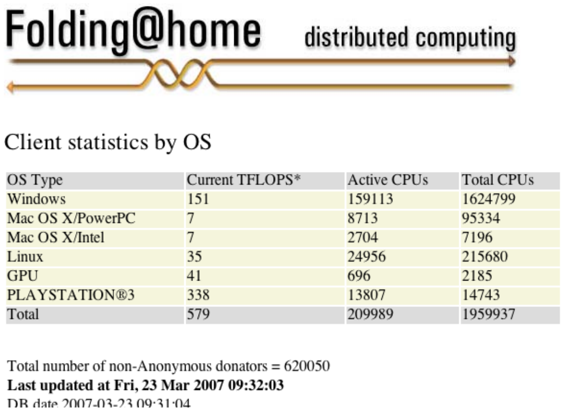


# My experience

## How one starts working on these problems

- Student in the CIHR BTP
- 4 month rotation with Dr. Hoos and Alena Shmygelska
- ...led into 1 year project
- ...led into a state-of-the-art algorithm

# Folding in the News



Folding@home distributed computing

Client statistics by OS

| OS Type | Current TFLOPS* | Active CPUs | Total CPUs |
|---|---|---|---|
| Windows | 151 | 159113 | 1624799 |
| Mac OS X/PowerPC | 7 | 8713 | 95334 |
| Mac OS X/Intel | 7 | 2704 | 7196 |
| Linux | 35 | 24956 | 215680 |
| GPU | 41 | 696 | 2185 |
| PLAYSTATION®3 | 338 | 13807 | 14743 |
| Total | 579 | 209989 | 1959937 |

Total number of non-Anonymous donators = 620050

**Last updated at Fri, 23 Mar 2007 09:32:03**

DB date 2007-03-23 09:31:04

*With PS3 now part of our network, we will be able to address questions previously considered impossible to tackle computationally, with the goal of finding cures to some of the world's most life-threatening diseases.*
*– Dr. Vijay Pande*

[a]

*2008 Update: 1533 TFLOPS*

[a] IBM BlueGene/L system at DOE's Lawrence Livermore National Laboratory (LLNL) performs at ~~280.6~~ 478.2 teraflops

# Closing Remarks

- Protein structure determination in the lab is costly (time/money)
- Existing techniques such as homology modeling can sometimes be reliable in practice
- We currently lack a great energy model for computational methods
- We use simplified models to understand the process better
- Protein folding is a hard problem (and still very much an active area)