# Lattice Models

So far - classify models by detail

| detail | type | properties |
|---|---|---|
| high | quantum mechanical | very physical |
| | atomistic | some approximations, mostly physical terms |
| low | coarse grain | crude functions, approximations, often non-physical terms |

Another important property

- continuous vs discrete

# Discrete

How to simulate weather / flow over an airplane wing..

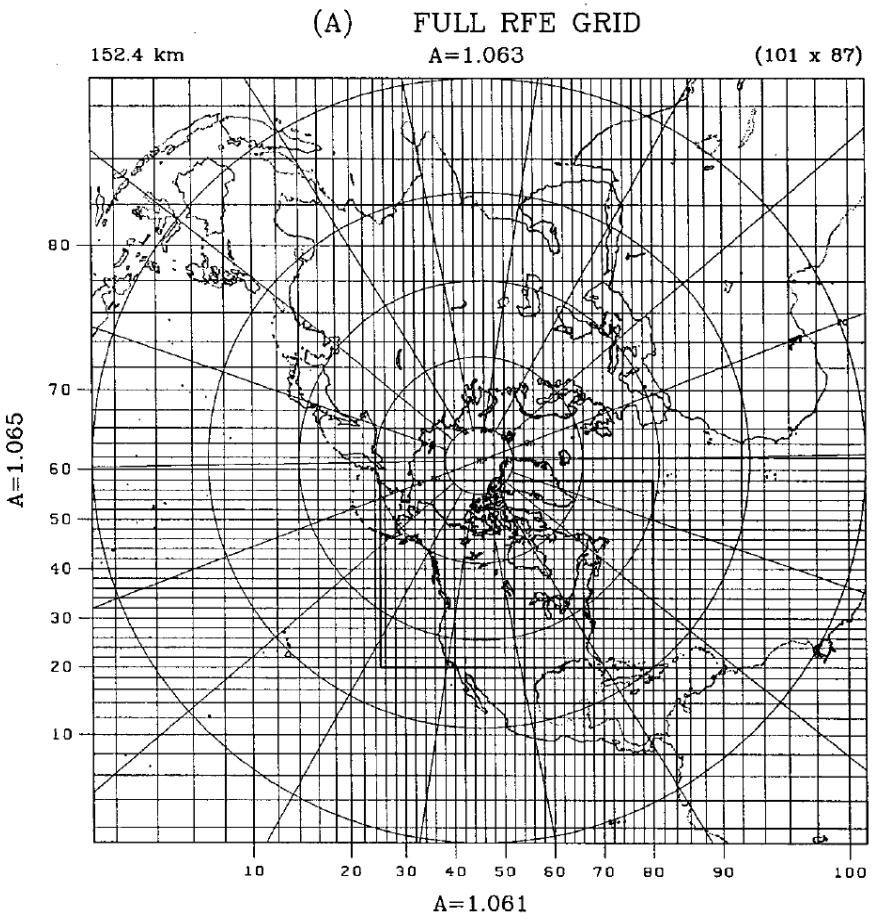
- take each atom
  - calculate interactions with neighbours, move system in time ? No

Make a grid

- store conditions at each grid point
- calculate interactions between grid points

Relevance to proteins ?

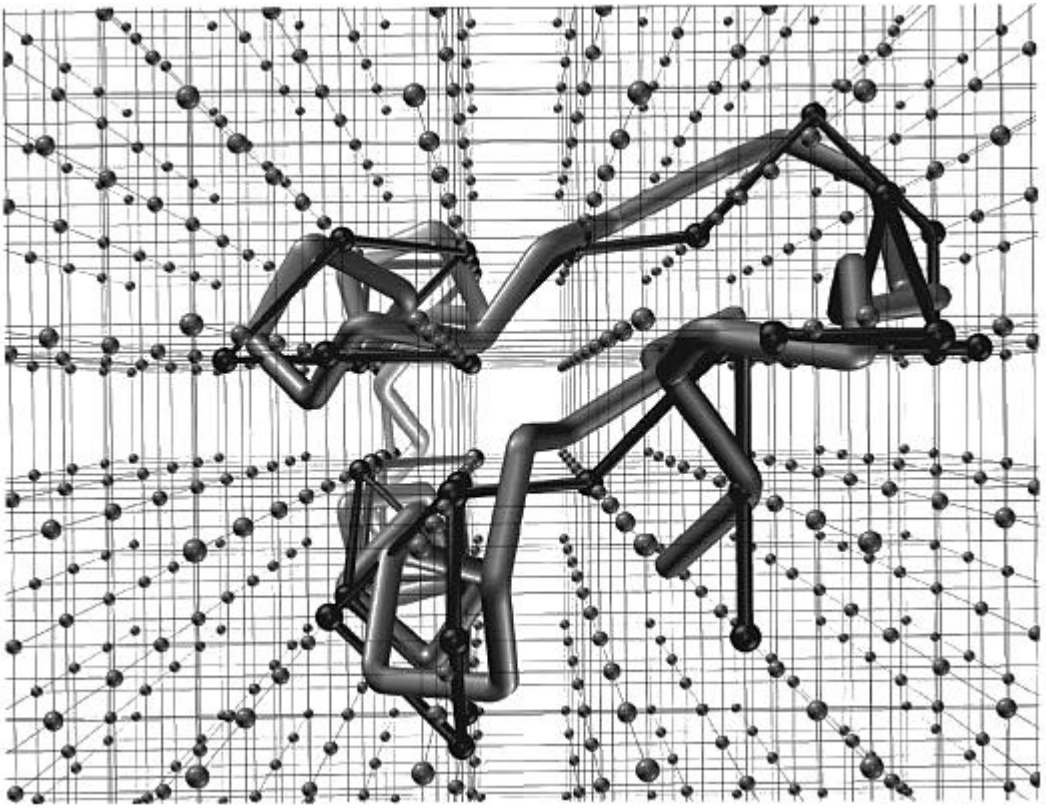
Discrete simulations ..

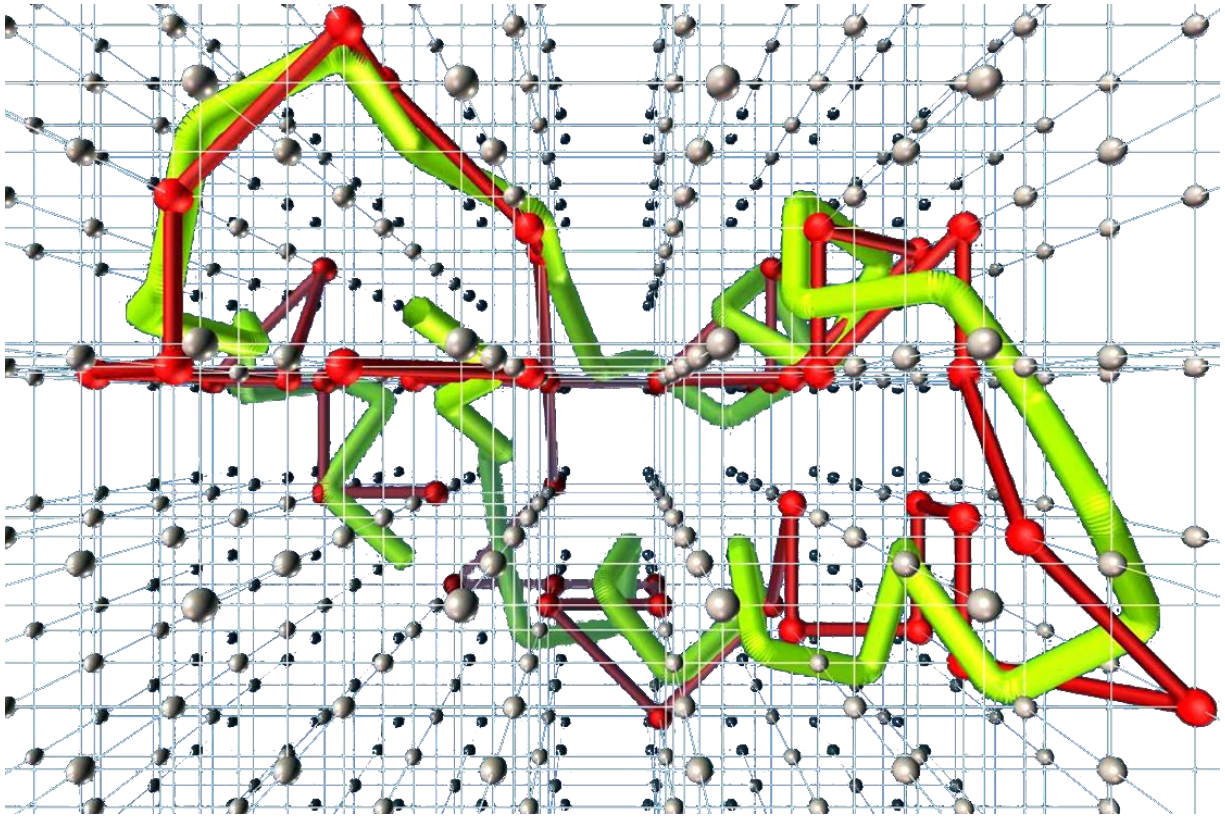



Benoit, R., Côté, J., Mailhot, J. Mon. Wea. Rev. 117, 1726-1750 (1989)

# Putting a protein on a lattice

Put atoms on nearest grid points





Reva, B.A., Finkelstein, A.V. Sanner, M., Olson, A.J., Skolnick, Prot Eng 10, 1123-1130 (1997)

# Continuous versus discrete

Continuous models

- coordinates (and other properties) take on any value
- typical properties
  - can take derivative with respect to coordinates
  - energy defined almost everywhere

Discrete

- coordinates (maybe more) are limited to certain values
  - think real/float versus integers
- examples
  - weather forecasts, oceanography, wind tunnels
  - finite element methods (engineering)
  - statistical mechanics (Ising model)

# Why ?

Do I want to model real proteins on a grid ? Not much

If I have a lattice

- Number of possibilities is much smaller (energies / structures)
- I can visit all / most of them

Big example in next lectures

- I can simulate evolutionary processes

Write this a bit more formally..

# Aim

Simulations so far

- long simulations necessary to sample conformational space
- to get average properties

$$\mathcal{A}_{obs} = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} \mathcal{A}_i \quad \text{or} \quad \mathcal{A}_{obs} = \frac{1}{b-a} \int_a^b \mathcal{A}(t)\, dt$$

With drastic simplifications either

1. increase $N_{obs}$ or
2. visit all possible (exhaustive enumeration) ..

# Exhaustive enumeration

- real world properties – average over all states
- probabilities depend on all states
- previously, we had "$N_{obs}$"

$$p_i = \frac{e^{\frac{-E_i}{kT}}}{Z} \text{ and } Z = \sum_{i}^{N_{states}} e^{\frac{-E_i}{kT}}$$

- in a simple system, $i$ can visit all $N_{states}$ states
  = exhaustive enumeration (examples soon)

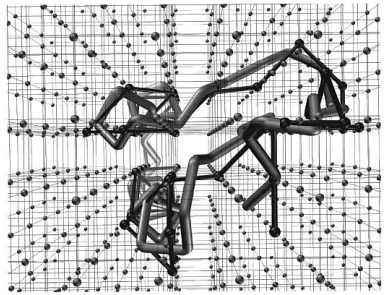
# Discrete proteins

How do we make proteins discrete ?

- most common
  - lattices, grids, (Gitter)

Are we modelling specific proteins ?

- sometimes



- usually not – more common
  - a very simple model to analyse some property

# Lattices, errors

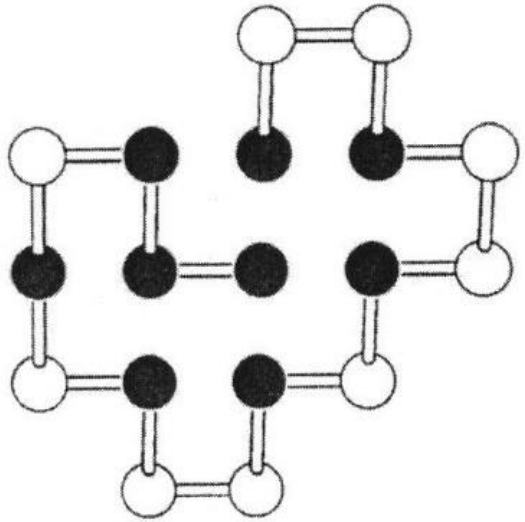
What would the error be on a lattice ?

- for 1 Å, should be ½ Å
- can be made arbitrarily small
- what if two continuous residues map to one point ?

Not the only (or best) criterion

- first, what would our energy look like ?

# Energy functions



Two philosophies

1. mimic approximation to real energies
    - earlier picture
2. simpler approach
    - continuous space for realistic simulations, real proteins
    - use simple model for some topic of interest

$$U = \sum_{i<j} c_{ij} \Delta(\vec{r_i}, \vec{r_j})$$

$c_{ij}$ is some parameterisation constant for types $i$ and $j$

$$\Delta(\vec{r_i}, \vec{r_j}) = \begin{cases} 1 & \text{if } i - j \neq 1 \text{ and } \left|\vec{r_i} - \vec{r_j}\right| = 1 \\ 0 & \text{otherwise} \end{cases}$$

diagram from Dill, Bromberg, Yue, Fiebig, Yee, Thomas and Chan, Protein Sci, 4, 561-602 (1995)

# Why simple energy functions ?

Simple functions (contact terms)

- some residues like to interact with each other
- will be happiest when the most favourable contacts are made (like a real protein)
- can reproduce very specific structures
  - interactions can be anything you want
- gross properties like hydrophobic packing

# Reduced alphabets

Typical question – we want to guess

- how does folding time depend on size ?
- how much hydrophobic area is exposed for some sequence ?

Do we need 20 amino acids ?

- general principle, consider 5 or 6 residue types
  - charged - (asp, glu)
  - charged + (lys, arg)
  - polar (thr, ser, gln, asn)
  - hydrophobic aromatic ( tyr, phe, his, trp)
  - hydrophobic aliphatic (ala, leu, val, ile, met, cys)
  - special (gly, pro)

# Reduced alphabets – HP model

History of protein structure

- most proteins have a hydrophobic core
- can this explain much of protein structure ?

Minimalist version

- two residue types (hydrophobic / polar, HP)

Say that protein structure is dominated by hydrophobic collapse

- two residue types are really enough for many calculations

- what properties can one reproduce with just
  - minimal geometry
  - hydrophobic / polar interactions

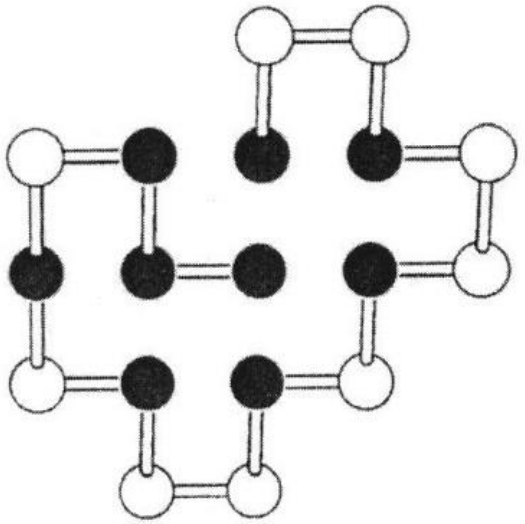
# Reduced dimensions

Do I care about specific real proteins ?

- not always

Is there a simple system which looks like a protein ?

- two dimensional protein
- very very simple protein ?
- 2-D, HP model

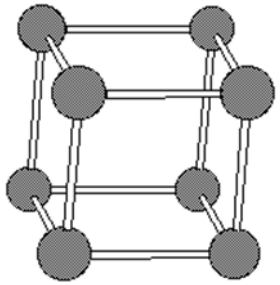


# models in these lectures

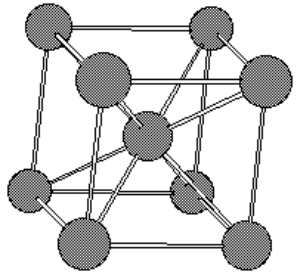
- mostly HP (hydrophobic / polar)
- sometimes 20 types of amino acid
- mostly square / cubic lattices

# Different types of lattice

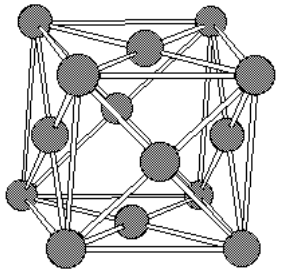
- simple cubic lattice



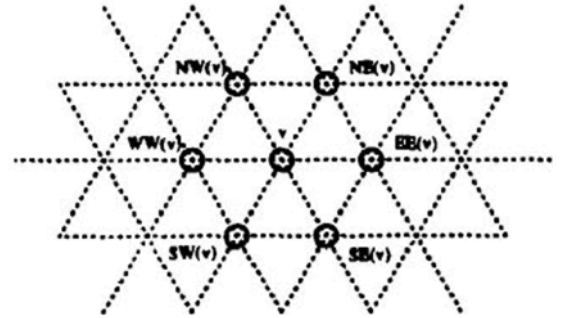
- body centred cubic



- face centred cubic



- triangular 2D / 3D




most important difference ?

- score functions count contacts
- how many neighbours does each have ?

triangular lattice from Agarwala et al, http://www.stanford.edu/class/cs374/Papers/p05-2.pdf
cubic lattice from http://wb.chem.lsu.edu/htdocs/people/sfwatkins/MERLOT/flattice/

# Why are lattice calculations so fast ?

- Normal code
  - for each particle
    - for each other particle
      - is it a neighbour ? calculate energy $O(n^2)$
- lattice code
  - for each particle
    - set up list of neighbour cells (often 6, 8, ..)
    - look if neighbour is occupied $O(n)$

What if we have a very realistic system ?

- all distances can be precalculated
  - 1 unit is 3.8 Å or 0.5 Å or …
- no more square roots $x^{½}$, cutoffs, …

# Calculations

We have some machinery, what kinds of calculations ?

- simulation (brief now, more later)
- others

Simulating on a lattice

- we do not have gradients of our energy terms (not much help if we do)
- we do know the energy of a configuration

Calls for Monte Carlo..

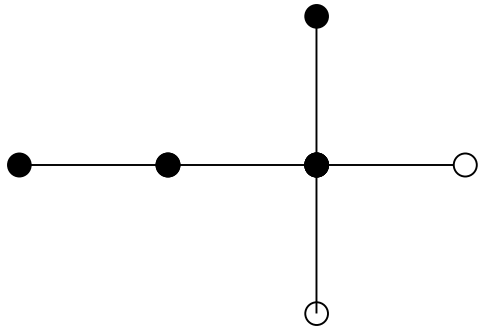
# Lattice simulations

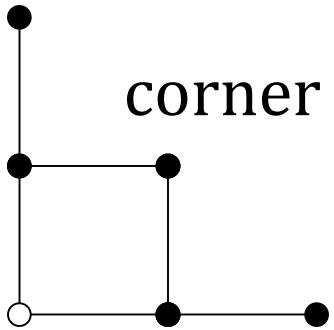
Monte Carlo - apply normal steps

- take a step
- calculate energy
  - accept / reject according to Metropolis criterion
- what would our moves look like ?
  - anything reasonable
  - from one starting point, should (eventually) be able to reach any other
  - want to be able to make big moves (speed – visiting conformations)
  - typical moves ..

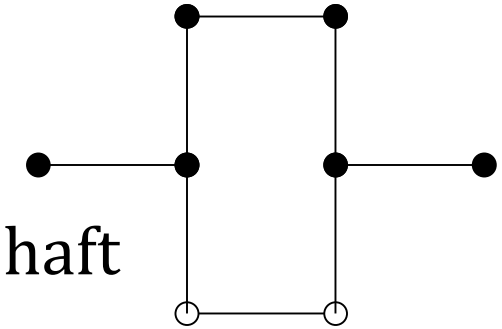
# Move sets

end move

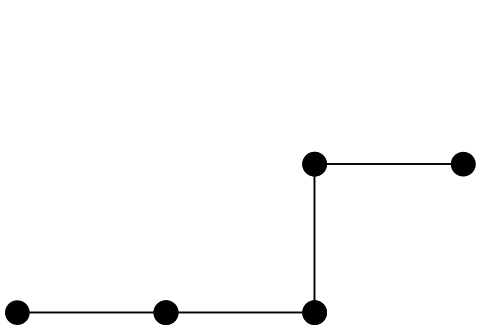


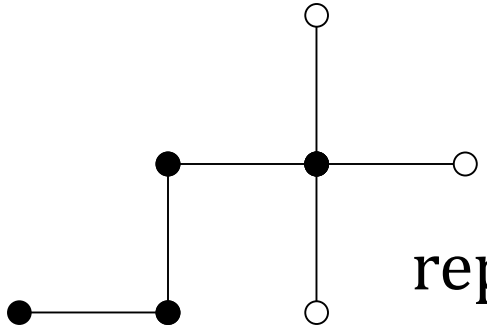


corner flip



crankshaft





reptation

# What can we get from simulating

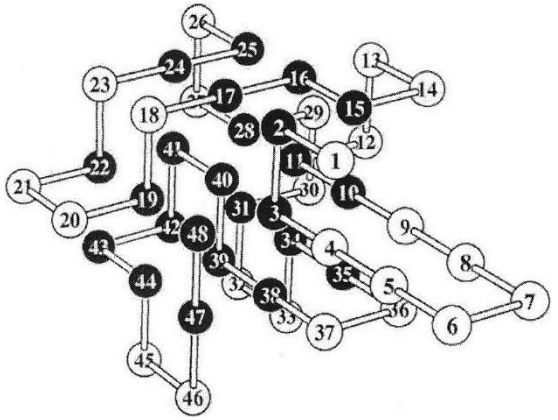
Take a system usually < 100 residues

- start from
  - random configurations
  - extended configurations
  - misfolded configurations
- run for $10^6$ or as many steps as you can afford
  - does a simulation always find a similar minimum energy ?
  - what is the energy spread of misfolded structures ?
  - are there many similar low energy structures ?
  - are there a large number of different low energy structures ?

# Results from simulations

From a 3D HP model, typical structure

- features ?
  - hydrophobic residues in middle



Compare with MD simulation

- biggest simulations in literature
  - small proteins
  - months of cpu time
  - do not find global minimum

More on simulations later…

diagram from Dill, … Chan, Protein Sci 4, 561-602 (1995), Principles of Protein Folding …

# Unique possibilities

Big problem with atomistic systems

- for any system more than about handful of residues
  - nearly impossible to visit all conformations
- for more than about 10 residues (maybe 15 or 20)
  - little evidence that the global minimum can be found

Lattices

- exhaustive enumeration (visit all possibilities)
  - configuration
  - sequence
- location of optimal structure

# Exhaustive enumeration of conformations

Why bother ?

- define almost all the stat mech properties of a system
- remember partition function
- summation over all conformations $Z = \sum_i e^{\frac{-E_i}{kT}}$

If we visit every $i$ we can find things like

- free energies
- distribution of energies

How many configurations are there ? 2D HP model

- 16 residues in 2D is no problem
- in 3D, about 3 × 3 × 3 feasible

| length | num configurations |
|---|---|
| 14 | 110 188 |
| 16 | 802 075 |

# Exhaustive enumeration of sequences

20 amino acids

- too hard

5 or 6 amino acids

- quite realistic, but difficult

HP model ?

- 16 residues is easy (65 536 sequences)
- with this machinery, what can we do ?

# Example question

Folding

- what are driving forces ? (hydrophobic collapse, HP)
- what is first to form (local or long range ?)
- how smooth is the folding pathway
- more later

Evolution

- more later

Do all protein sequences fold ?

Sequence vs structure space ?

# Do all proteins fold ?

If I take a random amino acid sequence, is it a protein ?

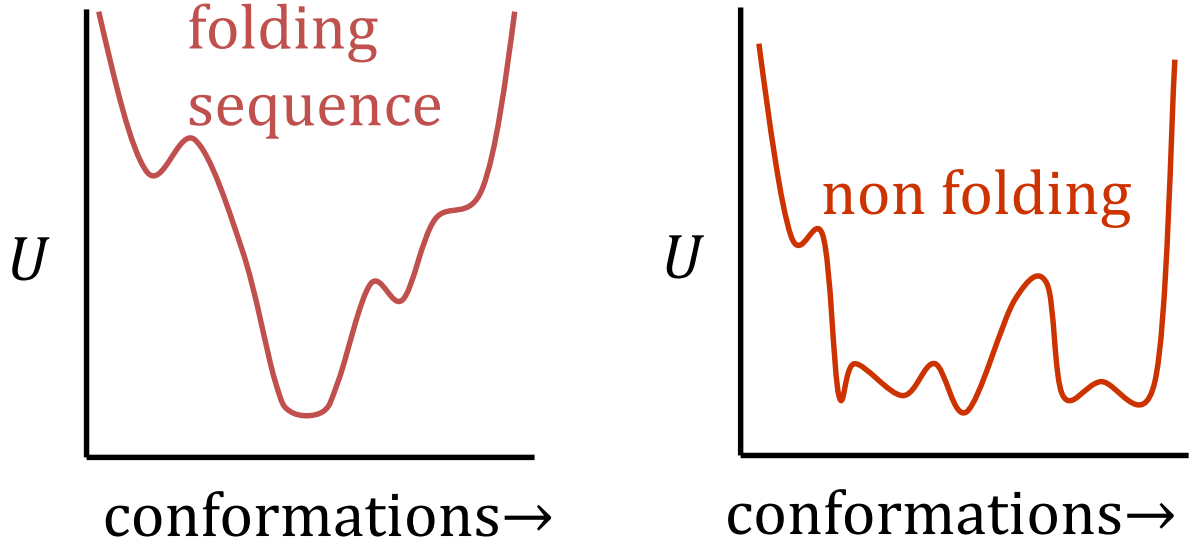
- experiment ? less than 1 in 100 fold
- test by MD simulation ?
  - cannot even fold one protein

Lattice models

- well studied problem

Definition

- important property
  - folding vs non-folding



# Folding versus non-folding

Non-folding in a lattice model

- find a sequence
  - visit all conformations
  - rank energies
- how many different conformations have the lowest energy ?
- how many have energy within $kT$ (could be visited at $T$) ?

Answers ?

- most random sequences do not fold
- intuitive example
  - a very very hydrophobic sequence is happy as long as it is compact
  - there are many ways to make it compact
- agrees with experiment

# Sequence versus structure space

From earlier lectures

- different sequences may fold to nearly same structure
  - large number of different sequences known for
    - globins, $\beta$-sandwiches, …
- different structures ? usually have unrelated sequences

Can we see this from MD ? No.

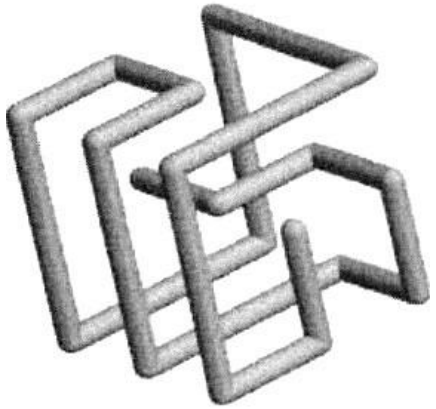
From lattice

- for each configuration
  - try every sequence and see if it is an energy minimum
- see how many sequences like each structure

# Favourite structures

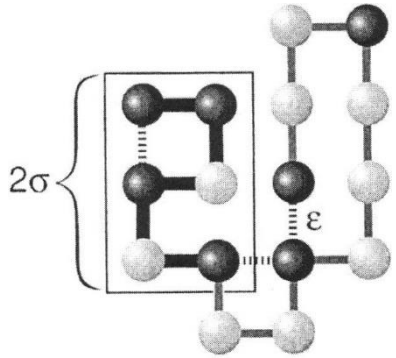
Some structures are the minimum energy for many sequences

- in a 3×3×3 HP model, there are 100's of sequences which like this structure
- some structures are popular, some much less so
- in principle, totally agrees with nature
  - exact numbers have no meaning



from Li, Tang & Wingreen, Proteins, 49, 403-412 (2002), "Designability of protein structures"

# Problems and limitation of lattices




Statistical mechanics are completely valid, but…

- loss of detail
  - resolution is obvious
  - interpreting in physical (or structural) requires faith
    - example of α-helix in 2D
  - whole structural properties may be lost
    - chirality ? chirality of a helix
- discretisation
  - energies and configurations are discrete
  - if a property depends on number of states, results will be model-dependent

diagram from Dill, … Chan, Protein Sci 4, 561-602 (1995), Principles of Protein Folding …

# Relating lattices to the real world

Simple models and reduced alphabets

- only trends are believable
- some trends can be tested
    - how do results change with 2 versus 3 amino acids ?

For detailed models,

- dependence on lattice type and resolution

# Artefacts

Susceptibility to artefacts ? Examples

- dependence on alphabet size
  - how popular is a structure may depend on alphabet size
  - in simple alphabets and energies, there are less foldable structures
    - more complicated models make lowest energy more unique
- properties depend on kind of lattice
  - extreme example !
    - a triangular lattice has more foldable structures

Are we finished with lattices ?

- long complicated story .. neutral evolution / lattices