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..

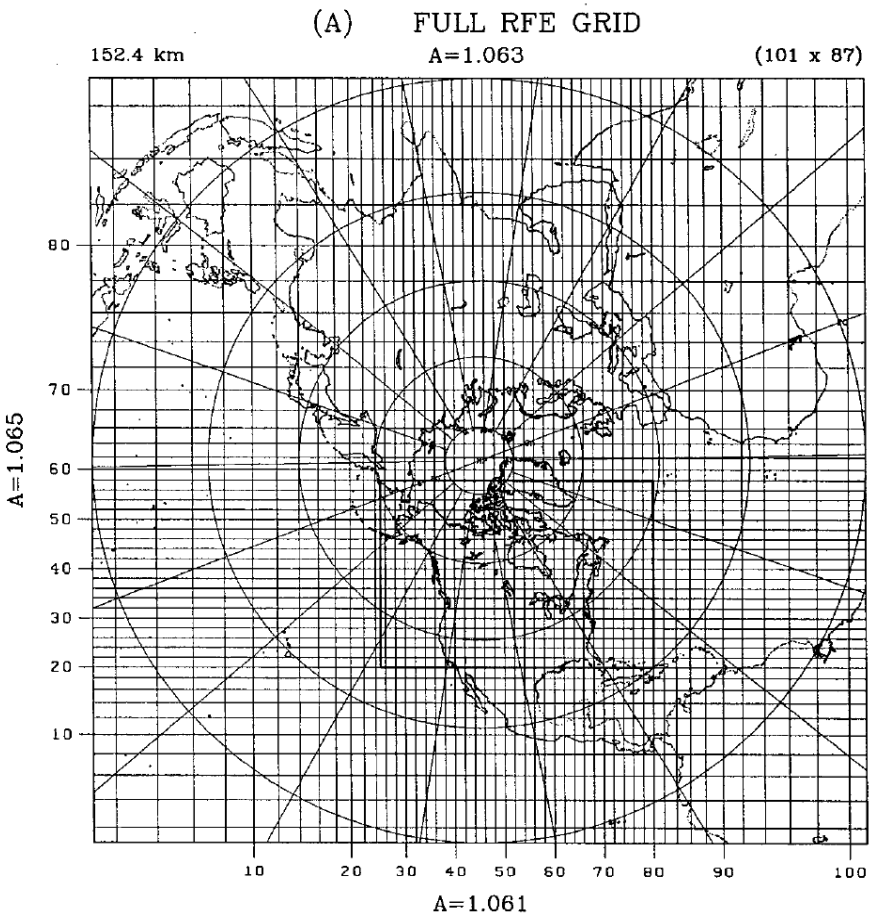
- take each atom
  - calculate interactions with neighbours, move system in time ? No

Make a grid

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

Relevance to proteins ?

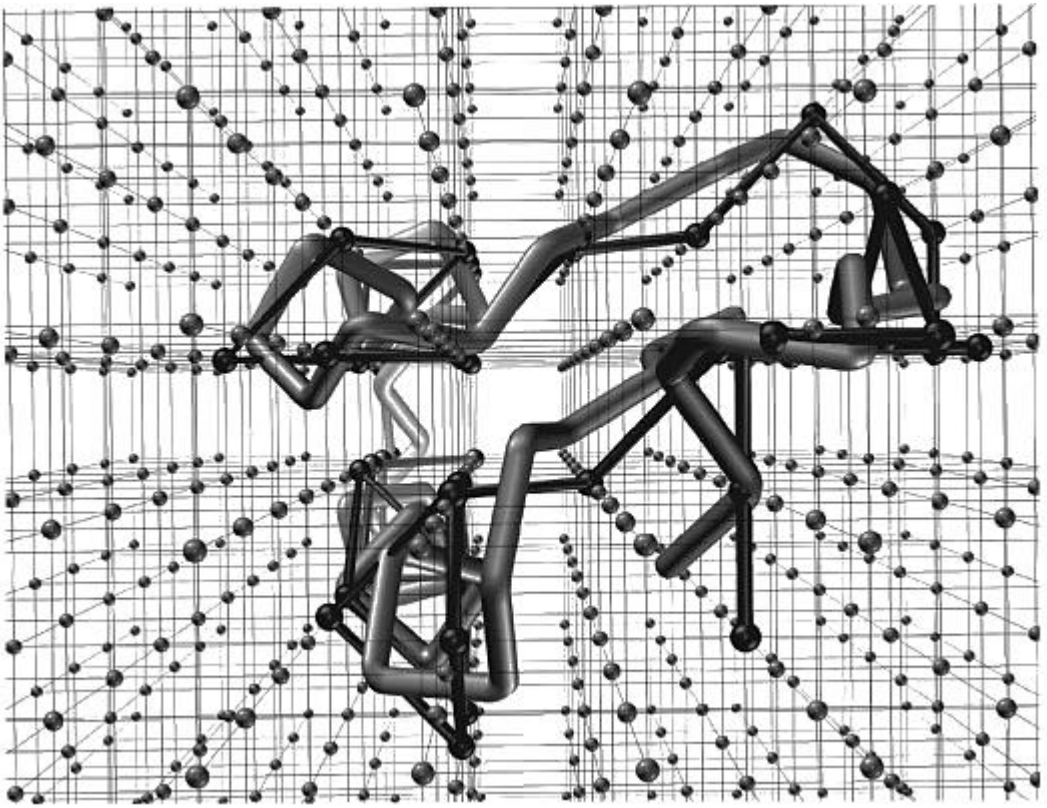
Discrete simulations ..

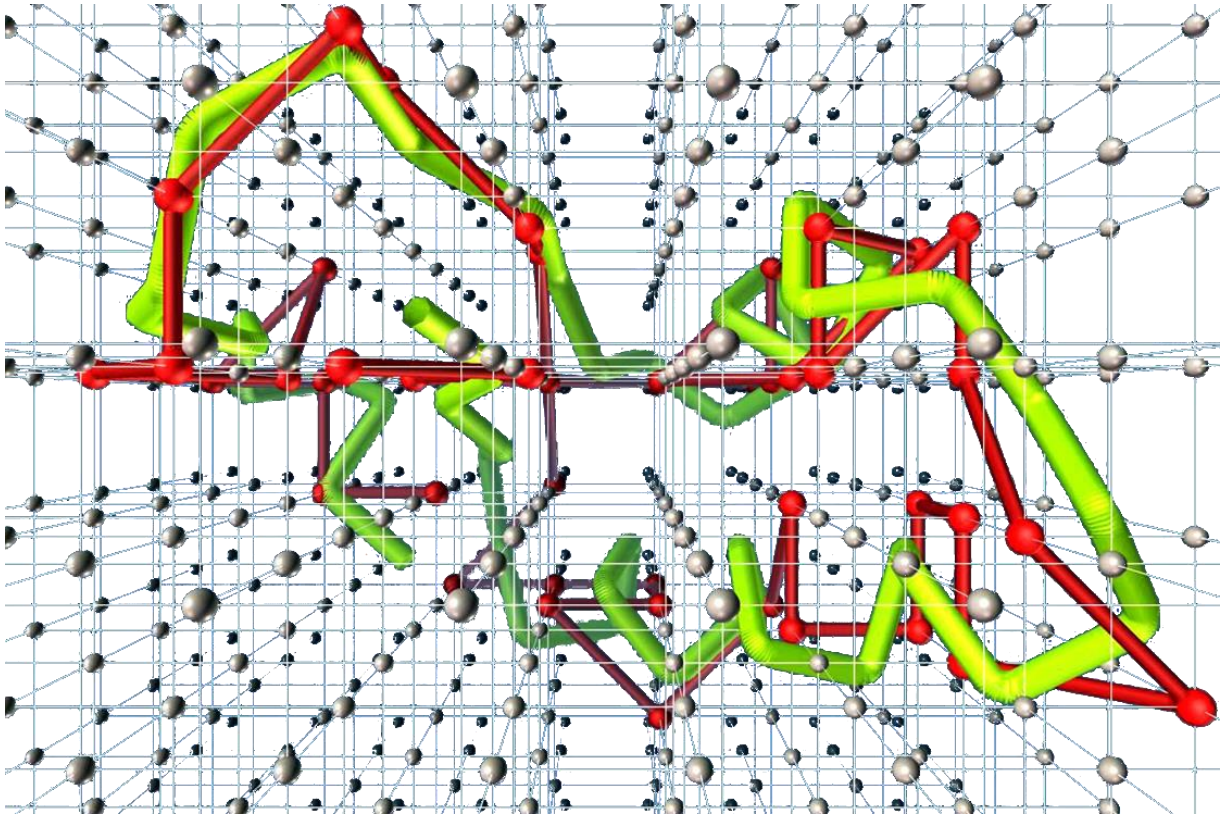


Benoit, R., Côté, J., Mailhot, J. Mon. Wea. Rev. 117, 1726-1750 (1989)

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 usually

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}"

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

Exhaustive enumeration

- in a simple system, one can visit all N_{states} states

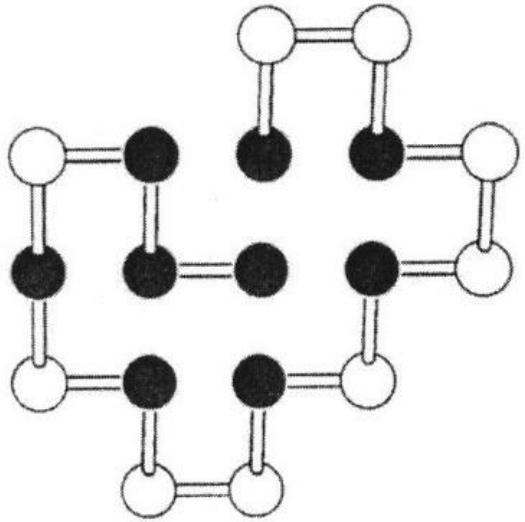
Discrete proteins

How do we make proteins discrete ?

- most common
  - lattices, grids, (Gitter)

- sometimes picture of real world
  - more common – a very simple model to analyse some property

Energy functions



Two philosophies

1. mimic approximation to real energies
   - earlier picture – not discussed here
2. simpler approach
   - 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 ?
- random sequences – how many fold to distinct conformations

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 enough for many calculations

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

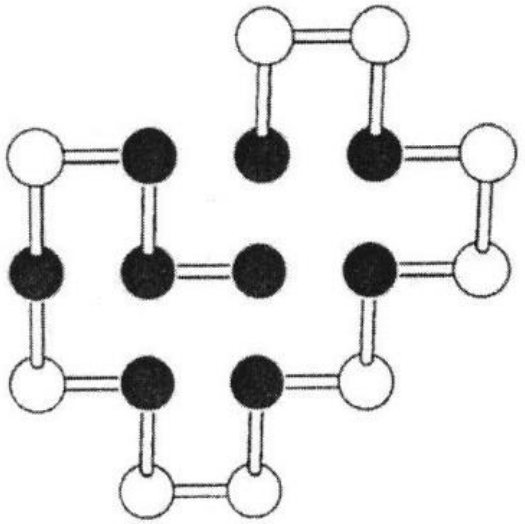
Reduced dimensions

Do I care about specific real proteins ?

- not always – interested in general properties of proteins

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

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, 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

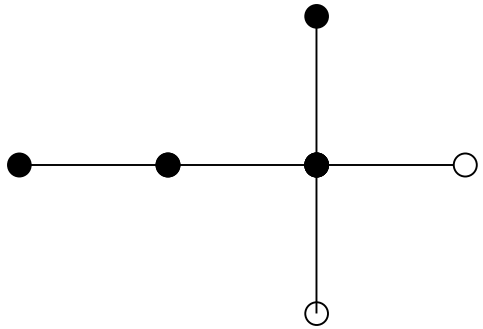
- take a step
- calculate energy (contact formulation)
    - accept / reject according to Metropolis criterion

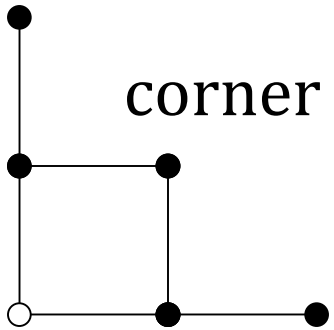
What would our moves look like ?

- 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 ..

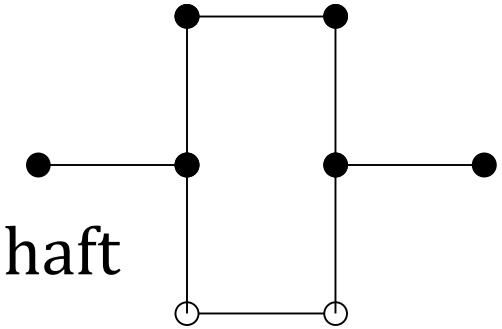
Move sets

end move

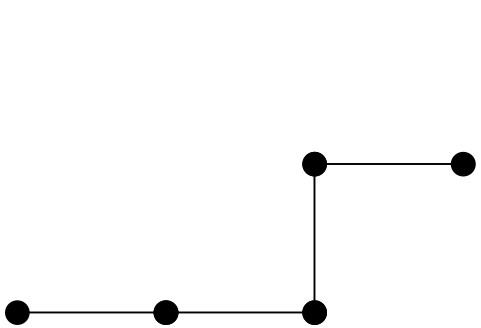


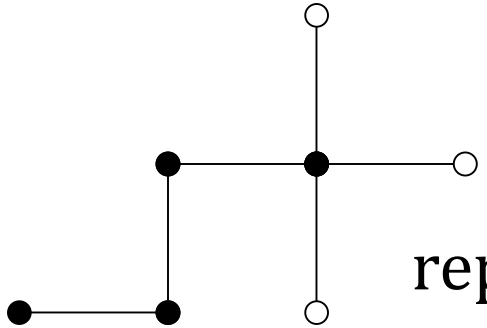


corner flip



crankshaft





reptation

What can we get from simulating ?

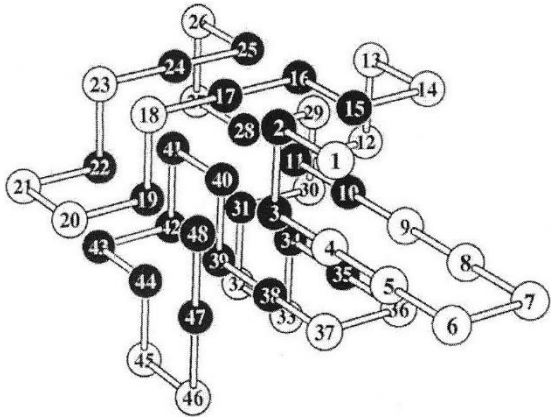
Take a system usually < 100 residues

- 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 $Z = \sum_i e^{\frac{-E_i}{kT}}$

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 (2^{16} = 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 ?

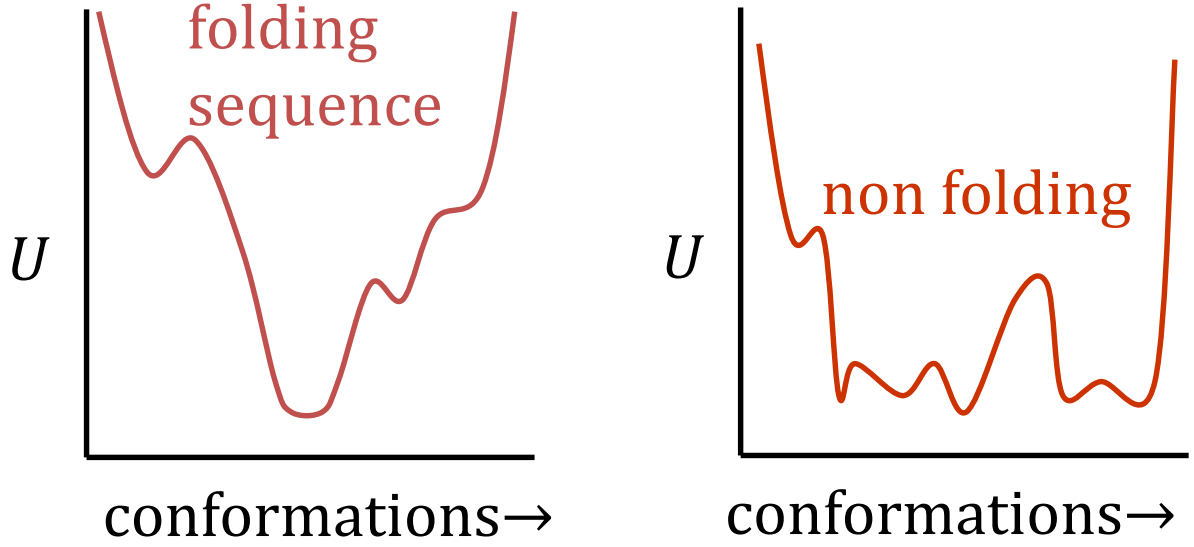
- experiment ? less than 1 in 100 fold
- test by MD simulation ?
  - cannot even fold one protein

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, β-sandwiches, …
- different structures ? usually have unrelated sequences

Can we see this from MD ? No.

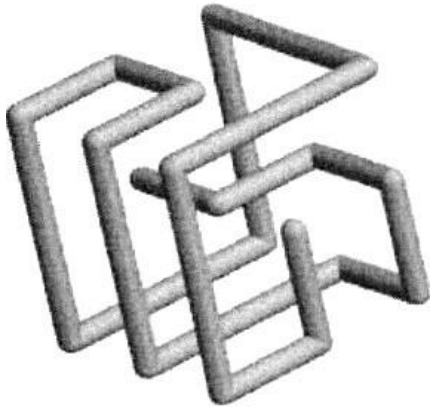
From lattice

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

Favourite structures

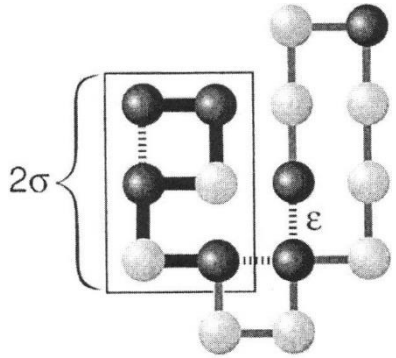
Some structures are the minimum energy for many sequences

- 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

Molecular Evolution

Why ?

- applications not possible with detailed models

Ingredients in this set of lectures

- models for proteins
  - simple representation - lattices, simple energy functions
- Boltzmann relation and partition function
  - ability to calculate probability of conformations

Aim

- from very few assumptions
- simulation which reproduces physical properties

Why lattice models ?

Earlier – building models

- how much detail - rather arbitrary

Here – minimal models

- one does not need serious chemistry to reproduce protein properties
- evolutionary pressure may not be real

Plan

Generalities
    - sources of evolutionary pressure
- example of unexpected evolutionary pressures (Darwinian)
- neutral networks
    - alternative explanation

Evolution observables

Phenotypes / population properties

- blue eyes, brown eyes (macroscopic)
- protein /nucleotide functions (molecular)

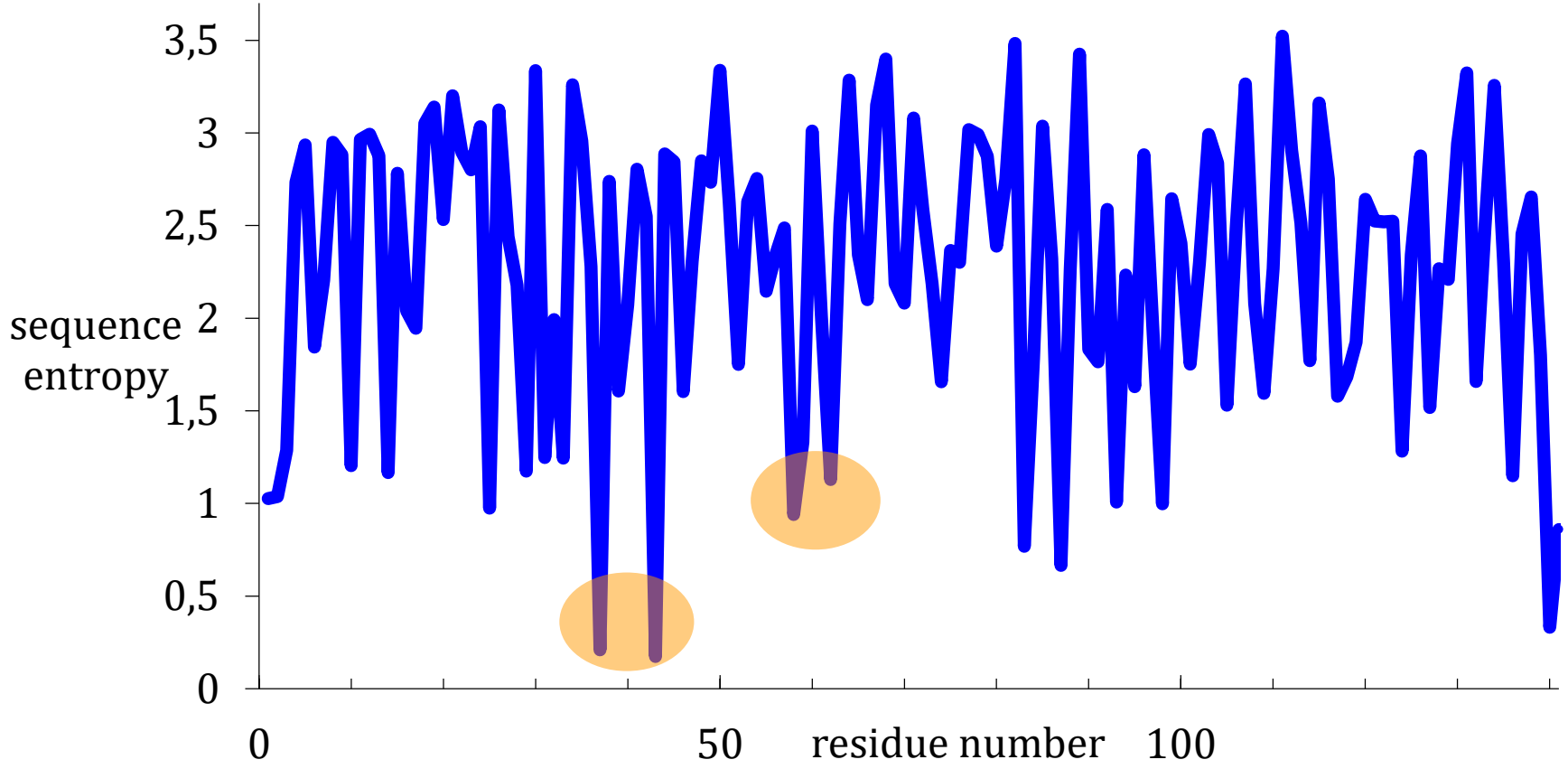
Consequence ?

- mostly look at evolution in terms of pressure on phenotypes
- classic adaptive Darwinism

First - a property to be explained later

Haemoglobin conservation

Look at residues 37, 43, 83 and 87



4 residues stand out as conserved

Sequence variability

Take family of related sequences

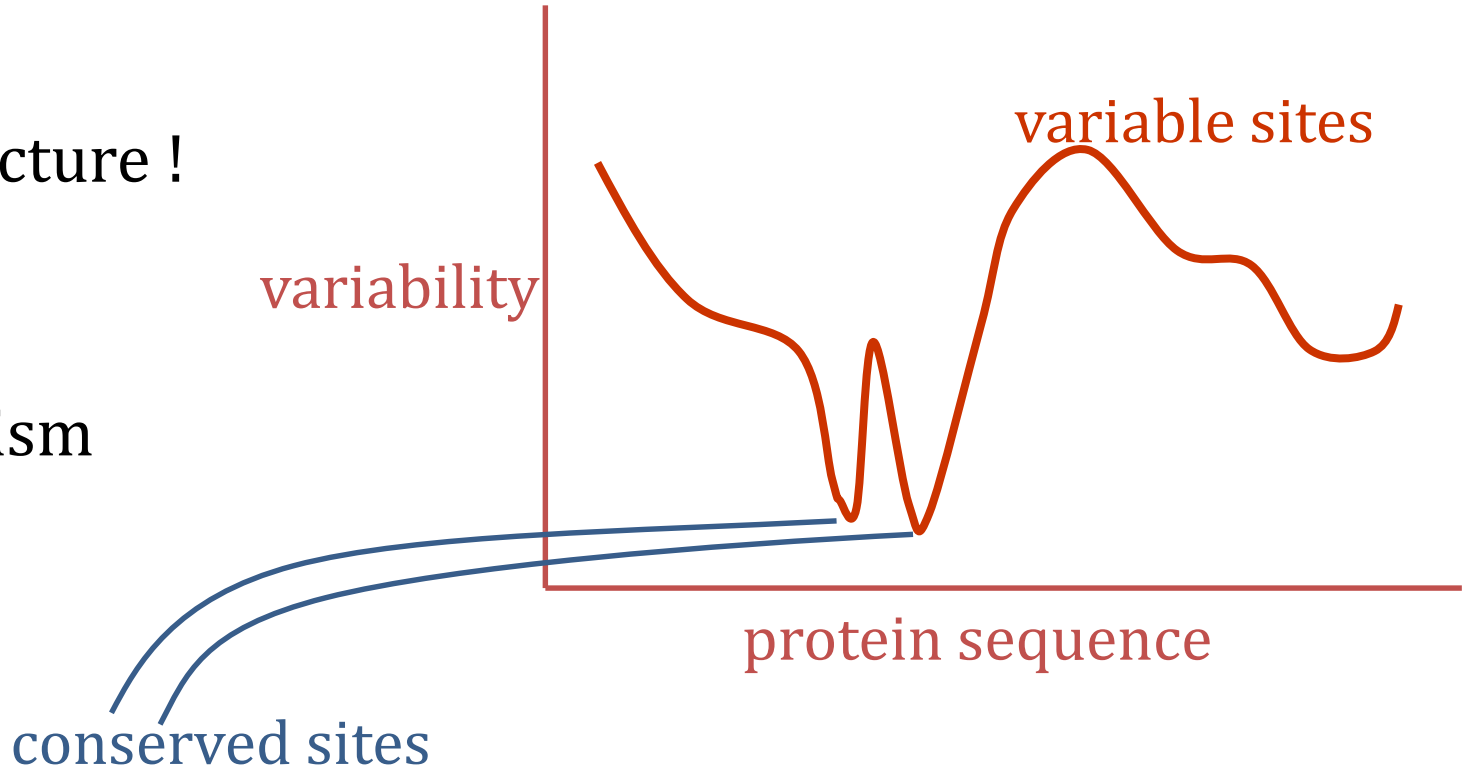
- see how conserved / variable they are

Variable sites

- are they unimportant ?

- remember this picture !

- return to Darwinism



Adaptive Darwinism

- I see a fish which lives behind a rock and eats seaweed
- A mouse is just the right size to squeeze through the hole in my wall
- Voltaire (1694-1778)

Master Pangloss taught ..

dass in dieser besten aller möglichen Welten, ...

> **„Es ist erwiesen“ sagte er, „dass die Dinge nicht anders sein können: denn da Alles zu einem Zweck geschaffen worden, ist Alles notwendigerweise zum denkbar besten Zweck in der Welt. Bemerken Sie wohl, dass die Nasen geschaffen wurden, um den Brillen als Unterlage zu dienen, und so tragen wir denn auch Brillen“**

Two aspects

- adaptation to glasses (evolution is directed)
- best of all possible worlds (we / the world are optimised)

Pangloss enseignait la métaphysico-théologo-cosmolonigologie

Classic Darwinism – molecular level

Obvious pressures

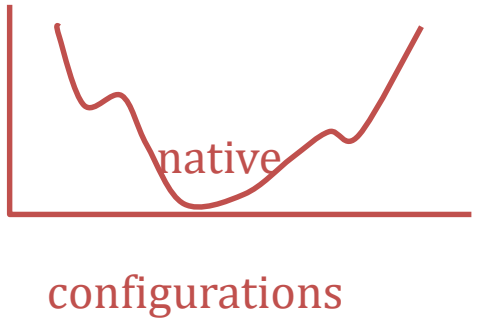
- function – protein must work
- stability – must be stable under your conditions

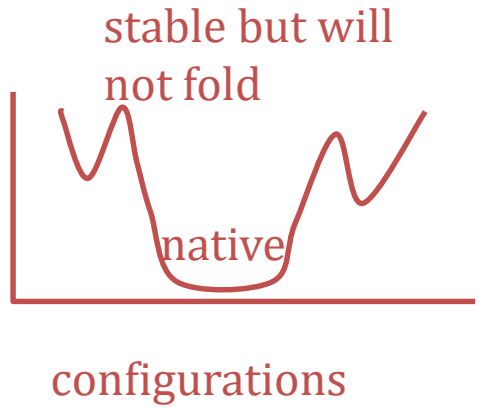
Less obvious, but simple

- folding – must fold in reasonable time

Less obvious, but reasonable

- mutation resistance





Other evolutionary pressures

Is it good to be resistant to mutation ?

- what if a gamma ray hits me and my children die ?
- more formally
  - a sequence (protein) is more likely to propagate if
    - it can be changed
    - it keeps functioning
- can this be modelled ?

Plan :

- be Darwinian
- (later) show why it is probabilistic (not Darwinian)

Simulating mutation resistance

Lattice simulations

- 25 residues, 2 dimensional, compact, 5×5 lattice
- 20 residue types (not two or 5 or 6)
- 1081 conformations
- remember we can calculate Z and stability
- for any sequence can say
  - will this sequence fold or not ? ΔG_{fold}
    - how different is lowest energy to other energies
- too big to check all sequences

Example calculation

- look at differences with and without evolution

*Taverna, DM and Goldstein RA, J. Mol. Biol. 315, 479-484 (2002) Why are proteins so robust to site mutations ?

Example evolution calculation

Evolution simulation

- apply mutations infrequently / randomly
- sequence must maintain
  - same structure
  - foldability
- for each member of population
  - check lowest energy configuration
    - if it has changed – sequence dies
  - check ΔG_{fold} based on Boltzmann probability of lowest energy structure
    - if sequence is not foldable – dies
  - of remaining sequences, randomly pick for reproduction

Simulation reminder

Simulations this semester

- system is not at equilibrium at start

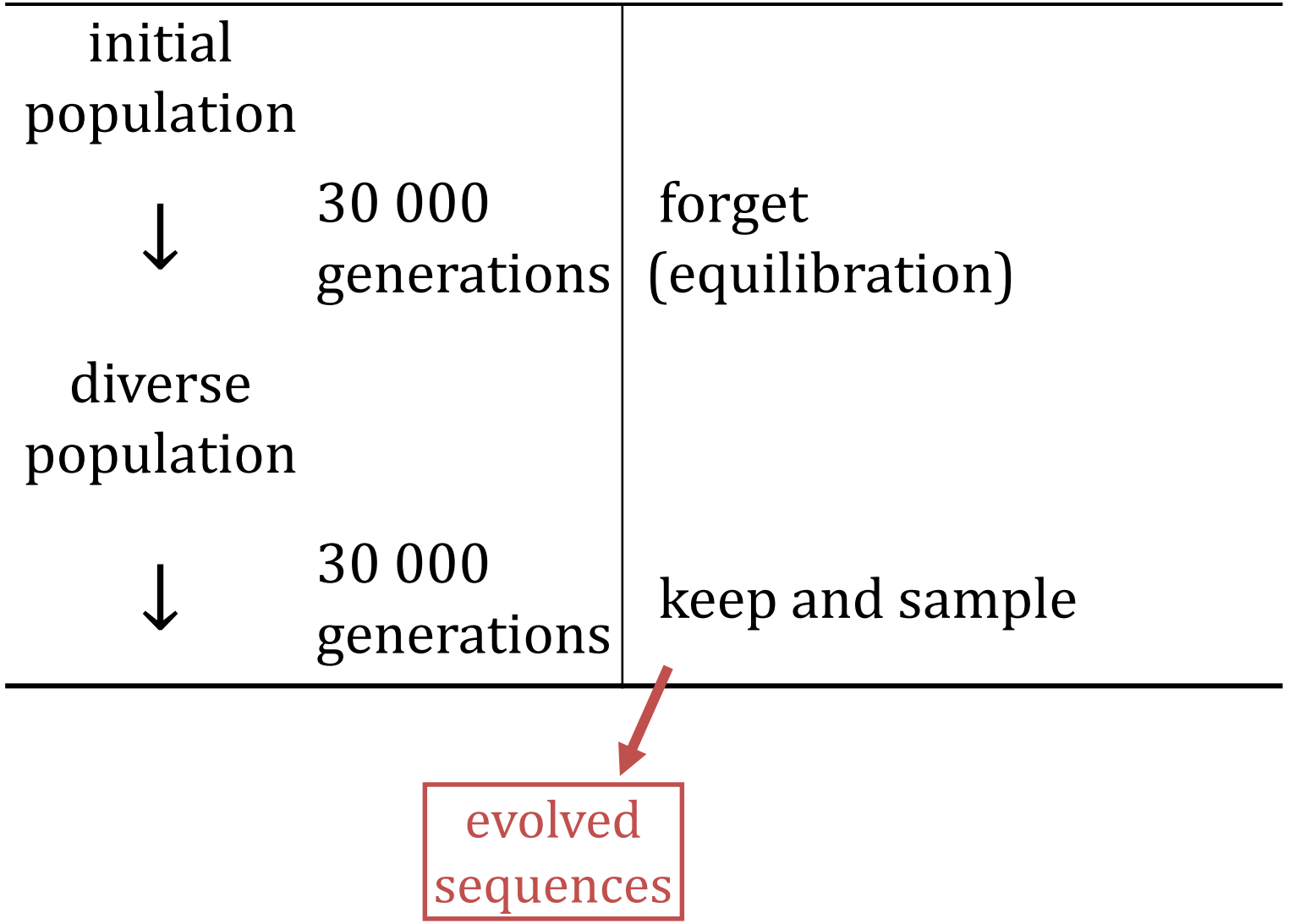
Normal procedure

- simulation for $n \times 1000$'s steps … throw away
- simulation more … keep for averaging and analysis

Comparing populations

Take a sequence which folds

- copy 3 000 times – initial population



Properties to look at

- How often does a mutation make a protein more stable ?
- How often does
  - a stable protein become more stable ? (not often)
  - an unstable protein become more stable ? (must be higher)
- Do the fractions differ between
  - random sequences (right hand side previous Folien)
  - evolved sequences (left hand side)

From simulation look at proteins with some ΔG (stability)

- after mutation get new ΔG
- look at large number of mutations, get probability $P(\Delta\,\Delta G > 0)$ of becoming even less stable

What do you expect ?

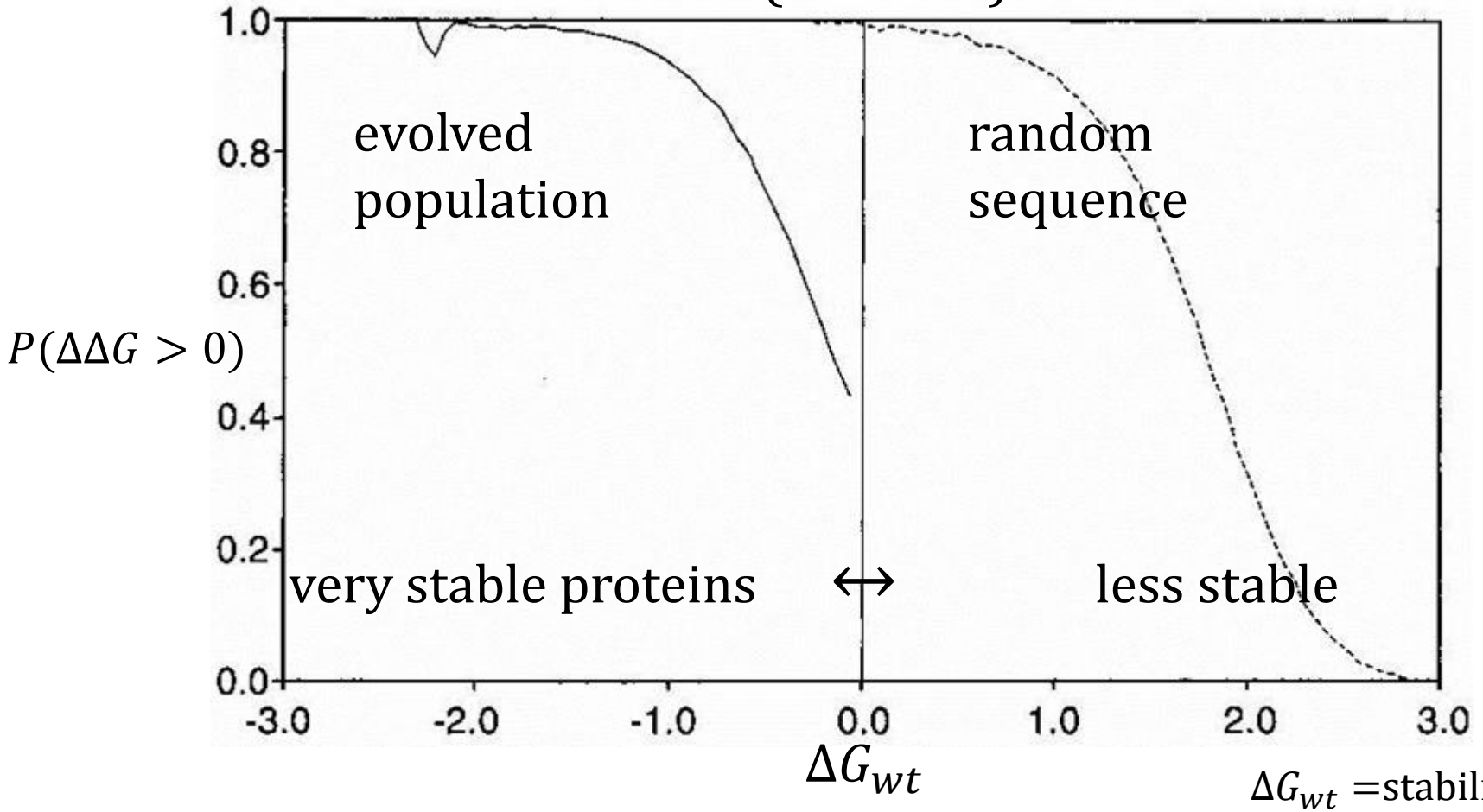
Evolved sequences must be more stable than random ones (obvious)

Will they also be more resistant to mutations ?

Simulation results

Take a sequence and have a look

- when it mutated and survived
  - how often did it become less stable $P(\Delta\,\Delta G > 0)$?

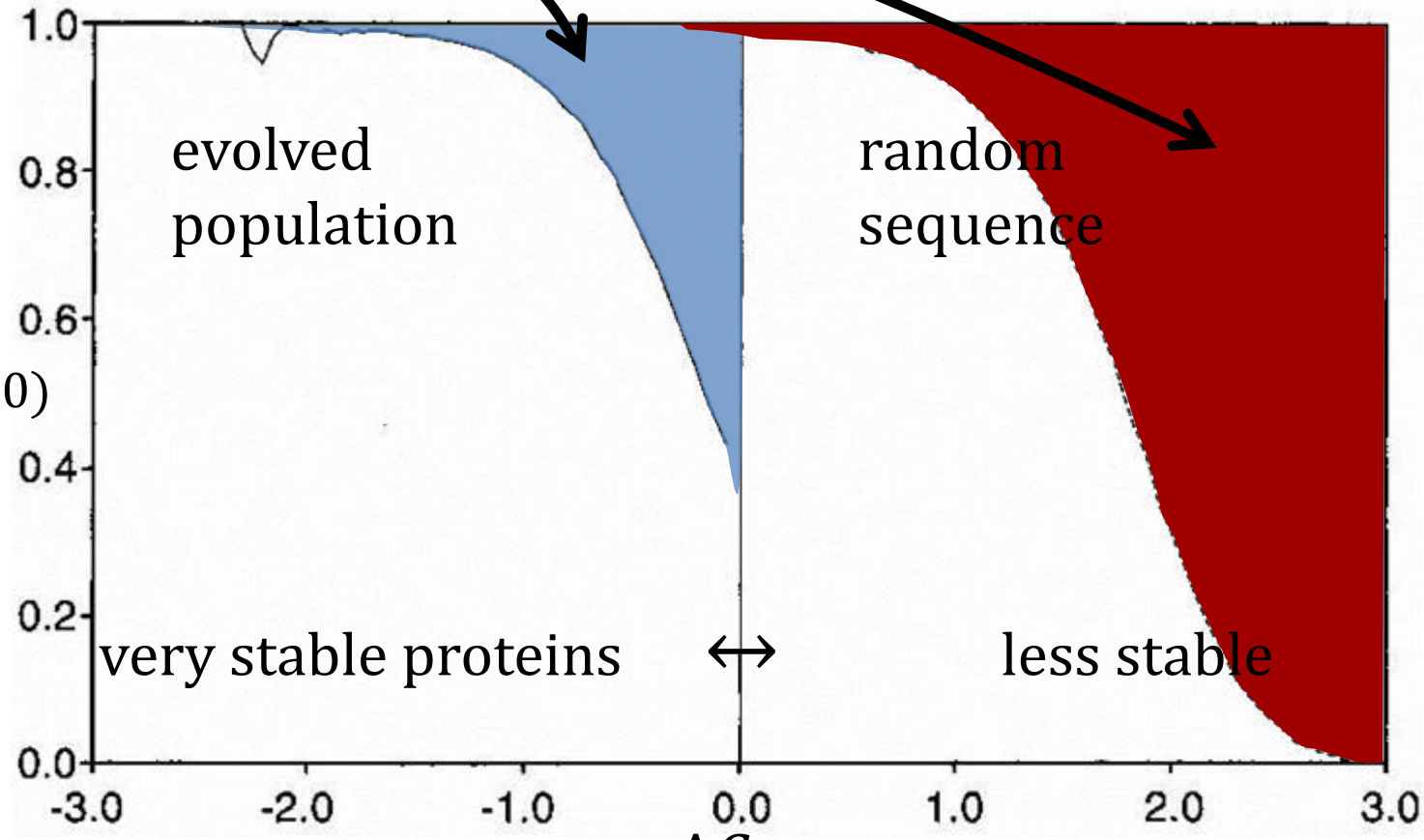


ΔG_{wt} =stability before mutation

Taverna, DM and Goldstein RA, J. Mol. Biol. 315, 479-484 (2002)

Simulation results

Becomes more stable



Taverna, DM and Goldstein RA, J. Mol. Biol. 315, 479-484 (2002)

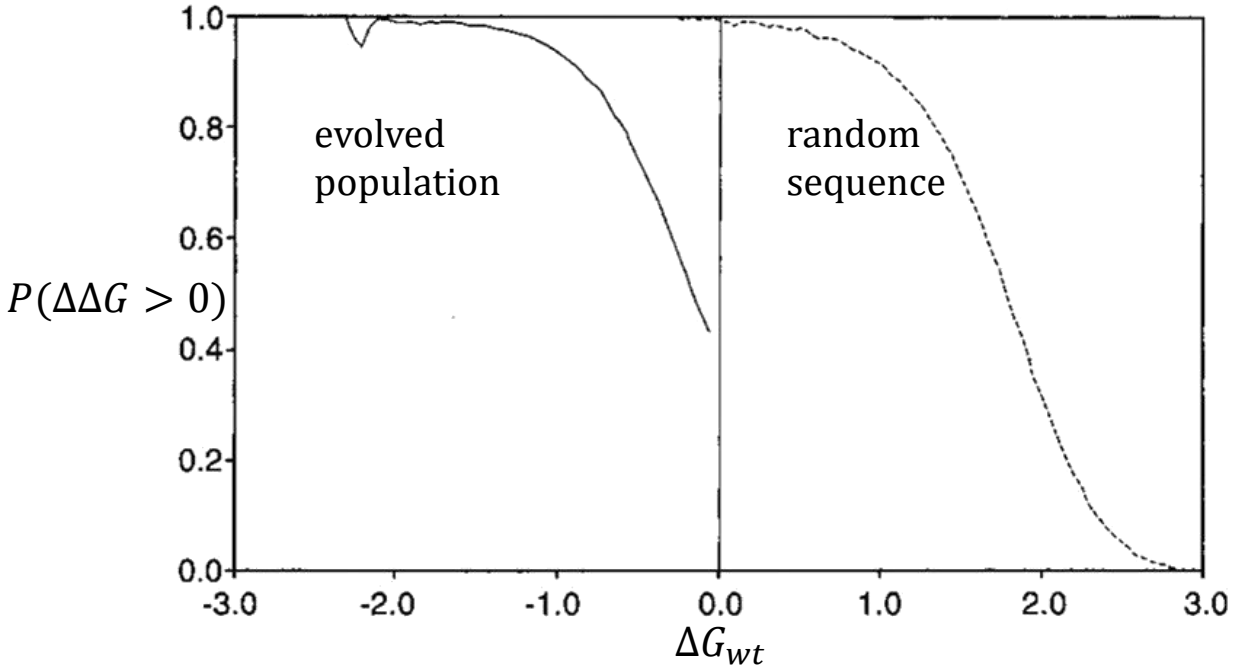
Interpreting results

random sequence

- unstable ($\Delta G > 0$)
  - not easy to make more stable
- stable ? ($\Delta G < 0$)
  - all mutations make it worse

evolved sequence

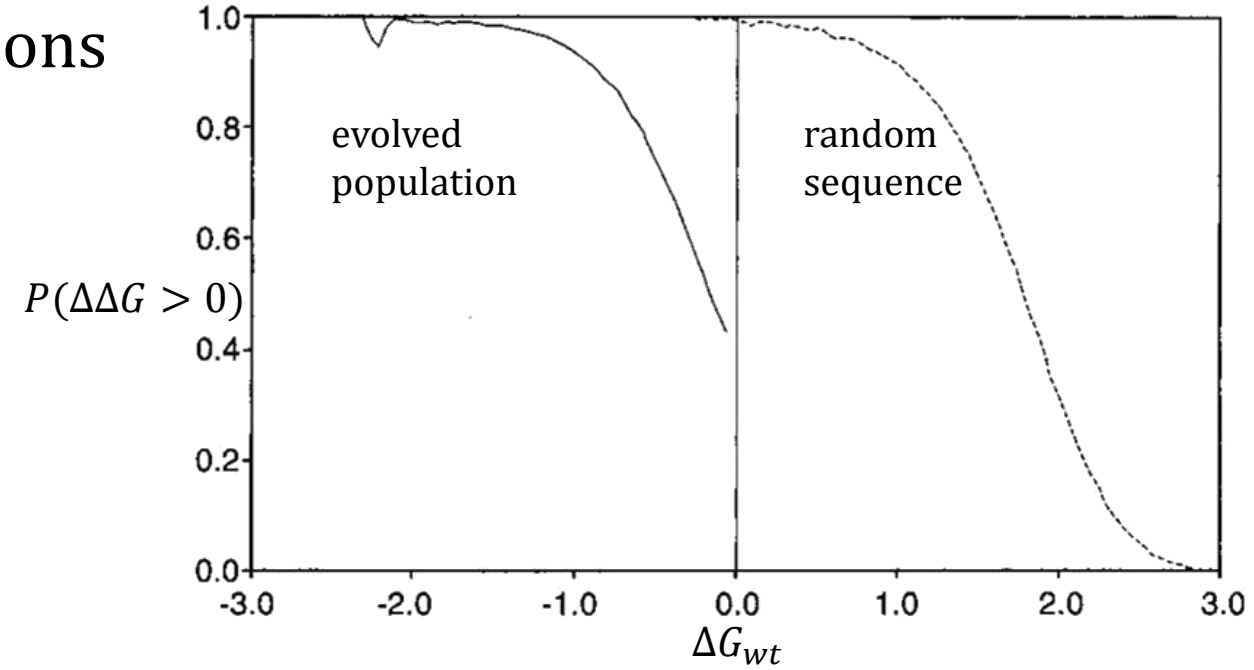
- very stable ?
  - cannot make better
- marginally stable ?
  - mutations often OK



Results explanation

Without explicitly adding idea

- evolution makes
  - more stable proteins (obvious)
  - proteins which survive mutations (why ?)



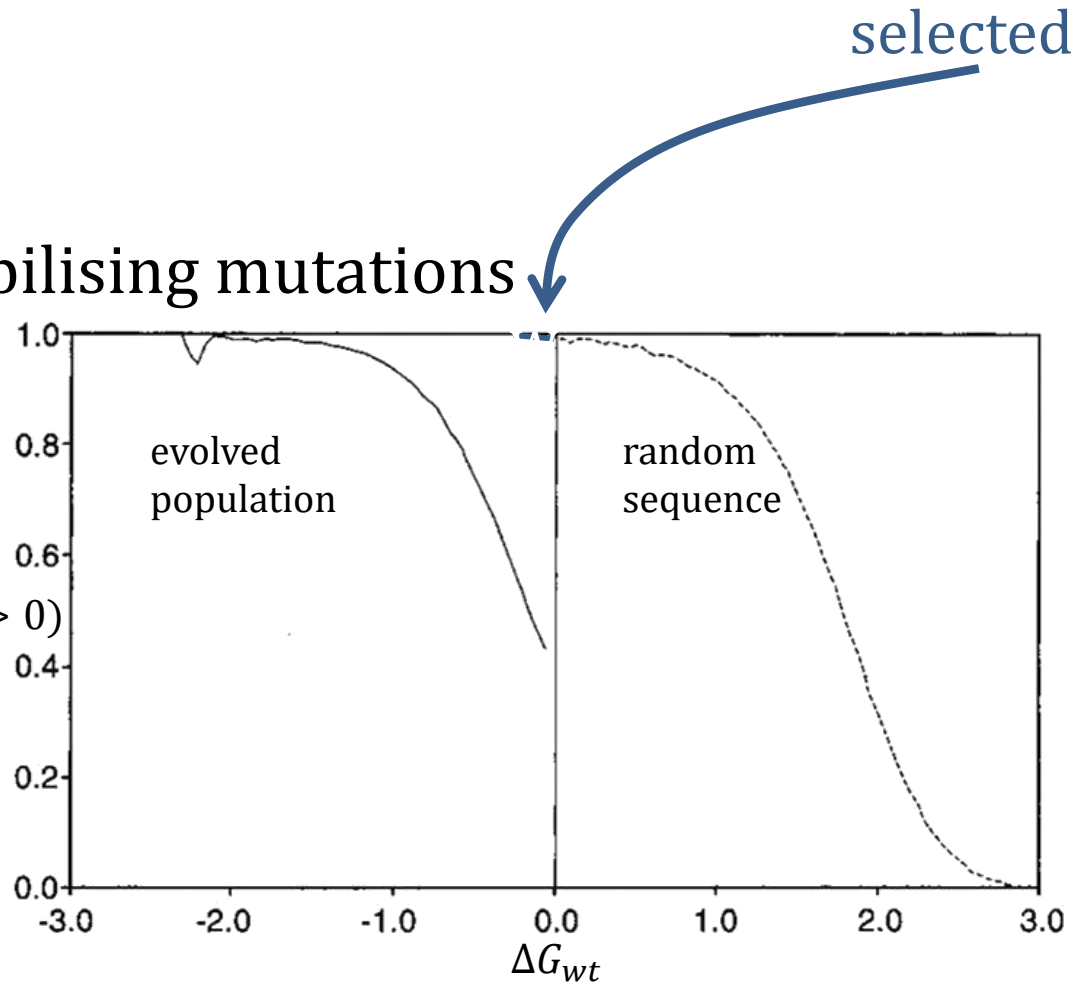
Agree with experiment ?

- small amount of the time
  - mutations have no effect
  - make protein more stable than natural version

Results explanation

Stability was selected

- moves population to left
- it should not change the fraction of stabilising mutations



Simulation selected for stable sequences

- of those stable sequences, did not select for mutation resistance
- $P(\Delta\Delta G)$ is a probability

- effect must come from somewhere else

Sequence variability interpretation

Typical part of sequence analysis

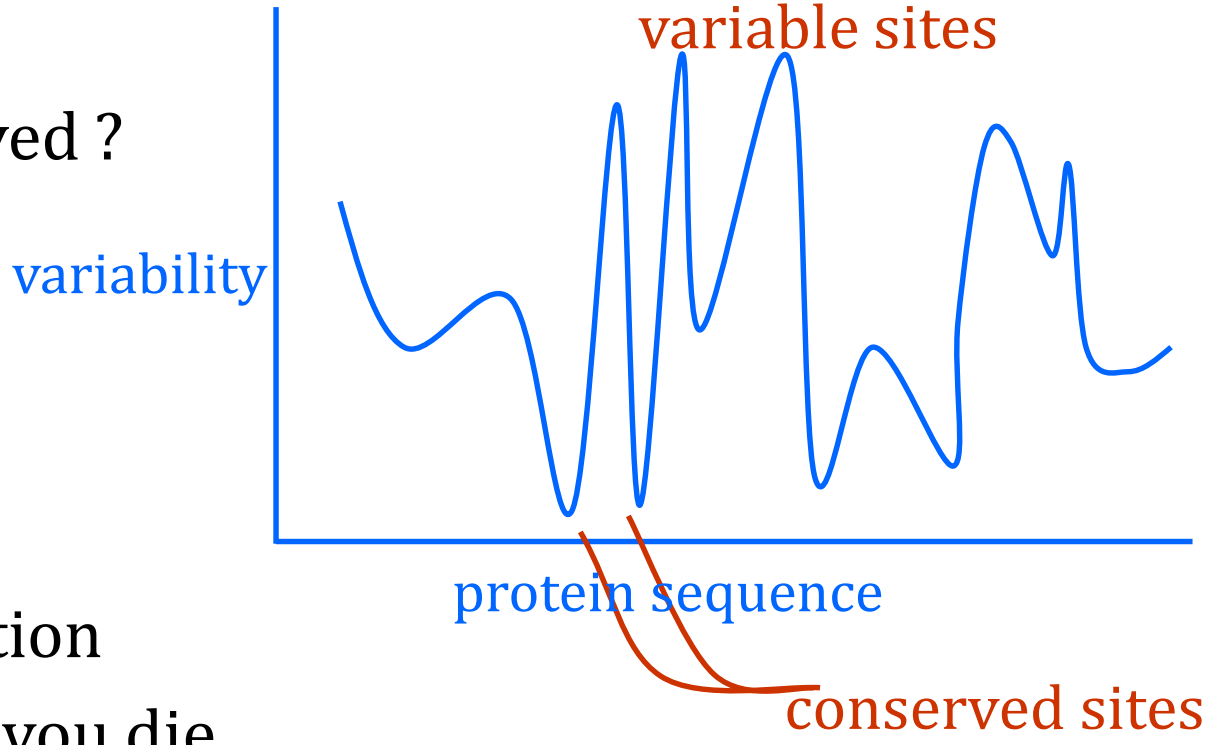
- look at collection of related sequences and see how conserved they are (conservation, profiles, sequence entropy, ..)

Why are some sites so well conserved ?

- function ?

Why do some sites vary ?

- old view: they do not matter
- this paper
  - this is a consequence of evolution
- if they are important and fragile, you die



Subtle evolutionary pressure ?

Is this an evolutionary pressure ?

- seems like a good idea to not die when mutated
- authors argue that the reason is different
- neutral evolution ...

so far

- very simple lattice model reproduces
- stability, evolutionary pressures

- not Darwin, but what is it ?

Simulating at the molecular level

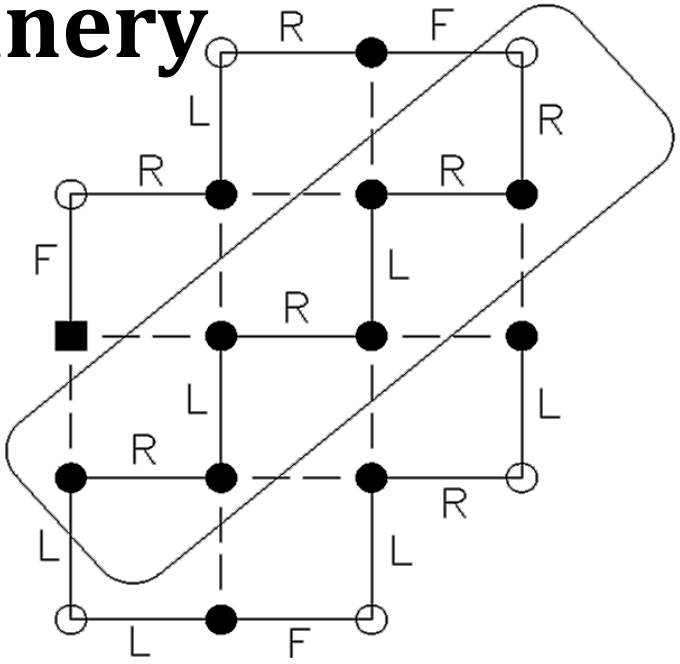
Basic idea

- take a population (maybe 10^3 or as big as possible)
  - make random changes
  - look at consequences
  - kill or reproduce molecules

Most popular

- RNA
  - for a given mutation, can guess at secondary structure
- Proteins
  - lots of lattice calculations

Simulation machinery



HP model in two dimensions

- length 18
  - one can look at all sequences
  - all conformations
  - ... for any sequence
    - can find minimum energy structure
  - for any structure
    - we can find all sequences which have this as minimum energy

Bornberg-Bauer, E (1997) Biophys J. 73, 2393-2403 How are model protein structures distributed in sequence space ?

what is a neutral mutation ?

General definition

- most mutations are a bad (deleterious) / few make you better
- some have no effect – neutral

In this model

- Sequence has a preferred ground state… after a mutation,
  - preferred conformation does not change – the mutation was neutral
- example
  - `HPHP`**`H`**`HH..` and `HPHP`**`P`**`HH..` have same ground state
  - this change does not cost anything in evolution
    - it is "neutral”

Calculations

Find popular structures

- which is best for many sequences
- collect these sequences
  - neutral set

Neutral mutations

- which of these sequences are connected by a point mutation?

Neutral mutations

Look at sites which can be changed

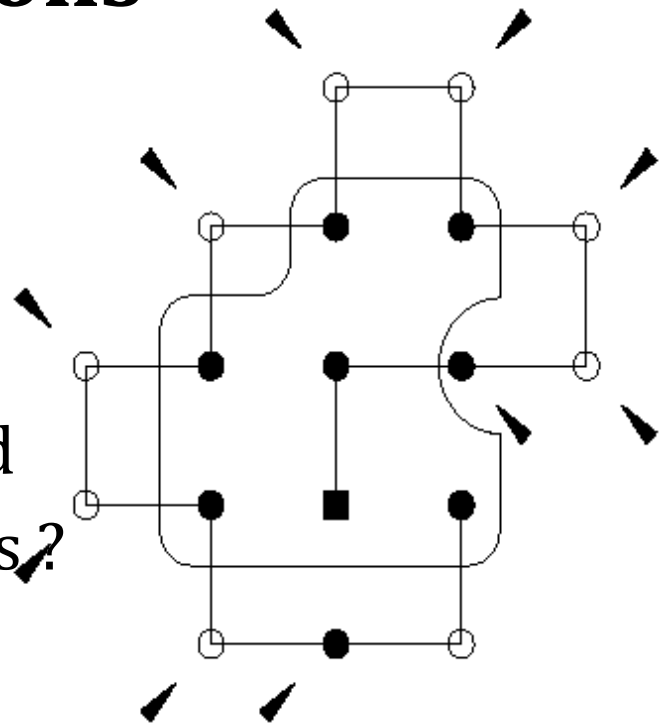
- many possible sequences

Can one mutate each to every other ?

- HPHP**HH**H.. and HPHP**PP**H are not connected

What can we say about the connected sequences ?

- form connected sets



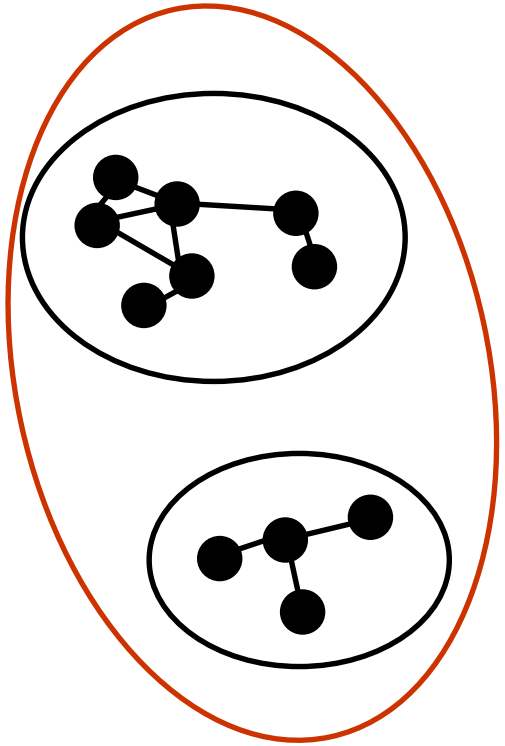
sites where neutral mutations were found

HPHP**HH**H and HPHP**PP**H may be a set, but not connected

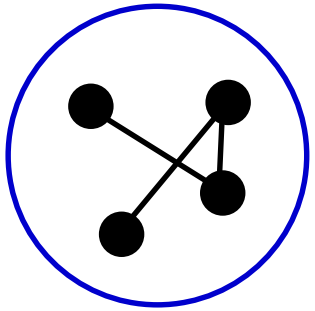
Bornberg-Bauer, E (1997) Biophys J. 73, 2393-2403 How are model protein structures distributed in sequence space ?

Connected and non-connected sets

Each dot is one protein sequence/structure



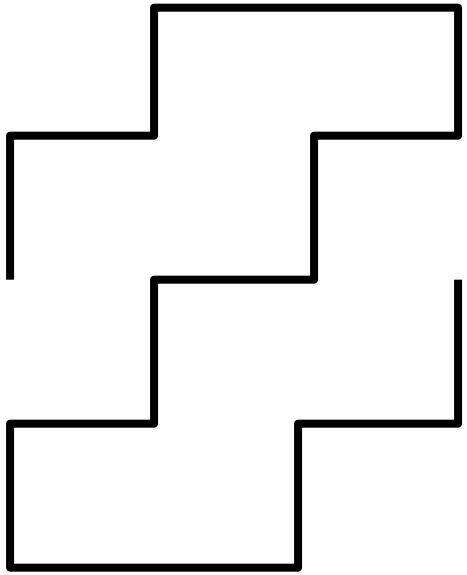
neutral set with two connected sets



neutral set and connected set

Neutral networks

Sequences which can turn into each other are "neutral network"



How big are the neutral sets ?

- about ¼ have more than 5 sequences
- most popular has 48 sequences
- lots of very rare structures

Are these sets fully connected ?
(can anyone eventually mutate into anyone else) ?

- about 80 % of time

Evolutionary consequences

- a population can quickly spread over a huge number of accessible sequences
- immense variation at molecular level is possible

- Can one hop between different connected networks ?
  - in this model – not so easily (≥ 2 mutations)

More interesting consequences

- some structures are hard to find by random moves
- some are very popular
- what does this say about mutation study ?

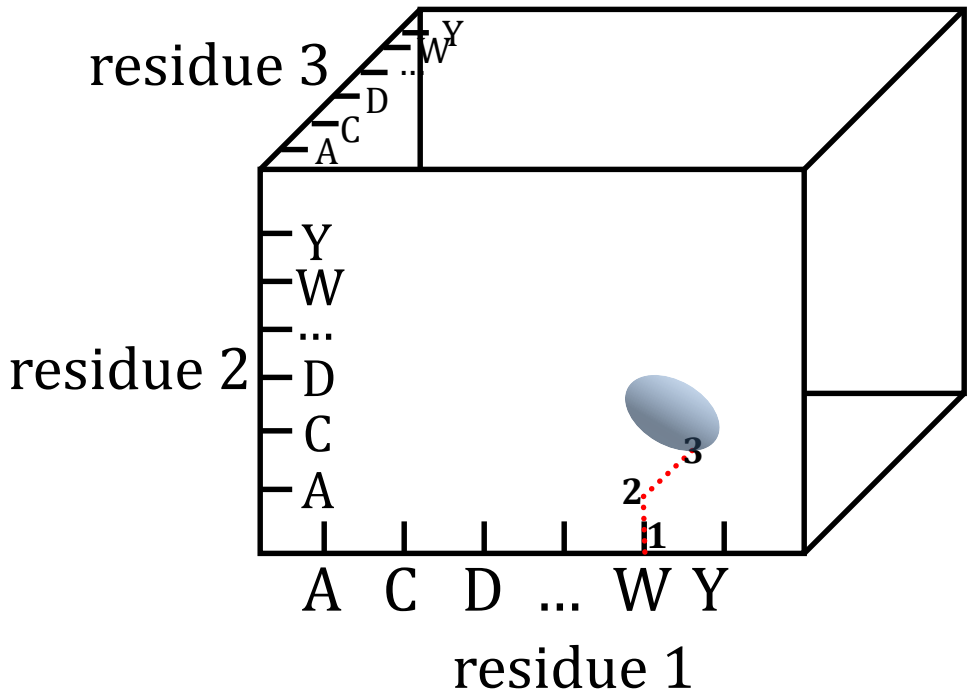
Mutation resistance revisited

Earlier slides

- it seems as if proteins evolve in order to be resistant to mutations (sounds Darwinian)

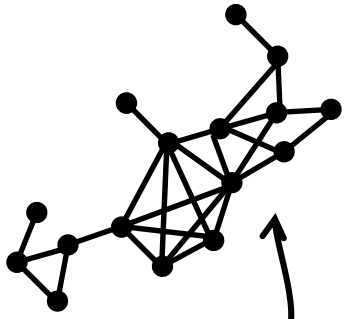
Alternative

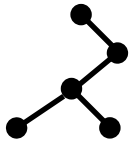
- think of sequence space
- a group of related sequences are a cluster in this space



Networks, probabilities, mutation resistance

huge network
1000's sequences





small network

mutate to here

- seems mutation resistant
  - lots of possibilities to mutate and maintain structure
  - more likely to be found (more sequences)
- mutate here ? likely to die

This is the alternative explanation of mutation resistance

- nothing to do with evolutionary pressure

Darwinian versus neutral evolution

Crux of these lectures

- Darwinian evolution – what you see is
  - most fit (selection pressure)
- Neutral evolution – what you see is
  - whatever is most likely to occur

Relevance to mutation resistance

- Darwinian
  - useful trait that will be selected for
- Neutral
  - larger neutral networks
    - by definition – mutation tolerant
    - because they are larger, more likely to be found

Summarise

- simple system lets you simulate long-term behaviour
- simulation selected for folding - found mutation resistance
- explanation comes from neutral networks

- not really an evolutionary trait