Molecular Evolution

Andrew Torda summer semester 2018, Struktur & Simulation

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)

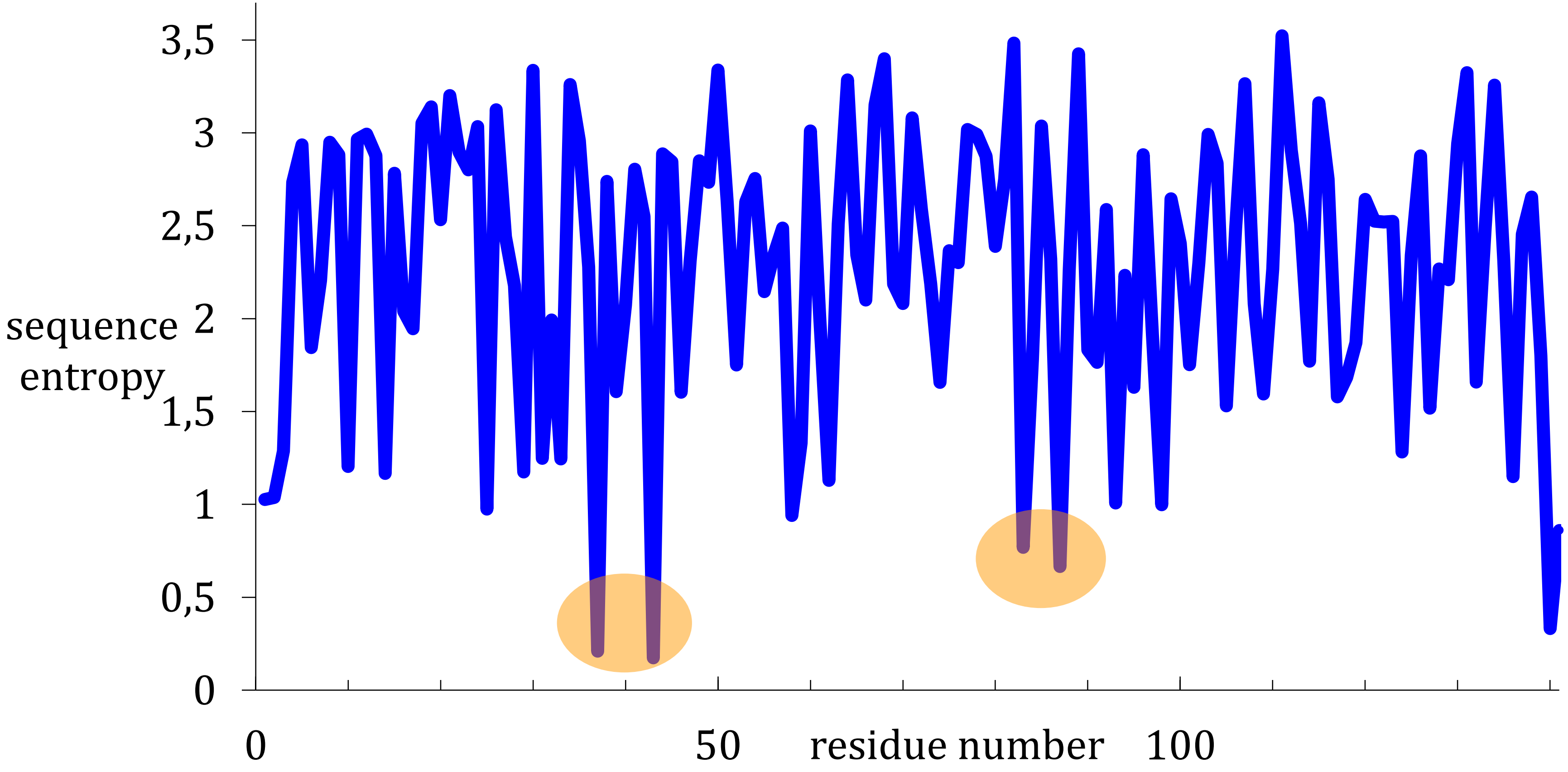
Consequence ?

- 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

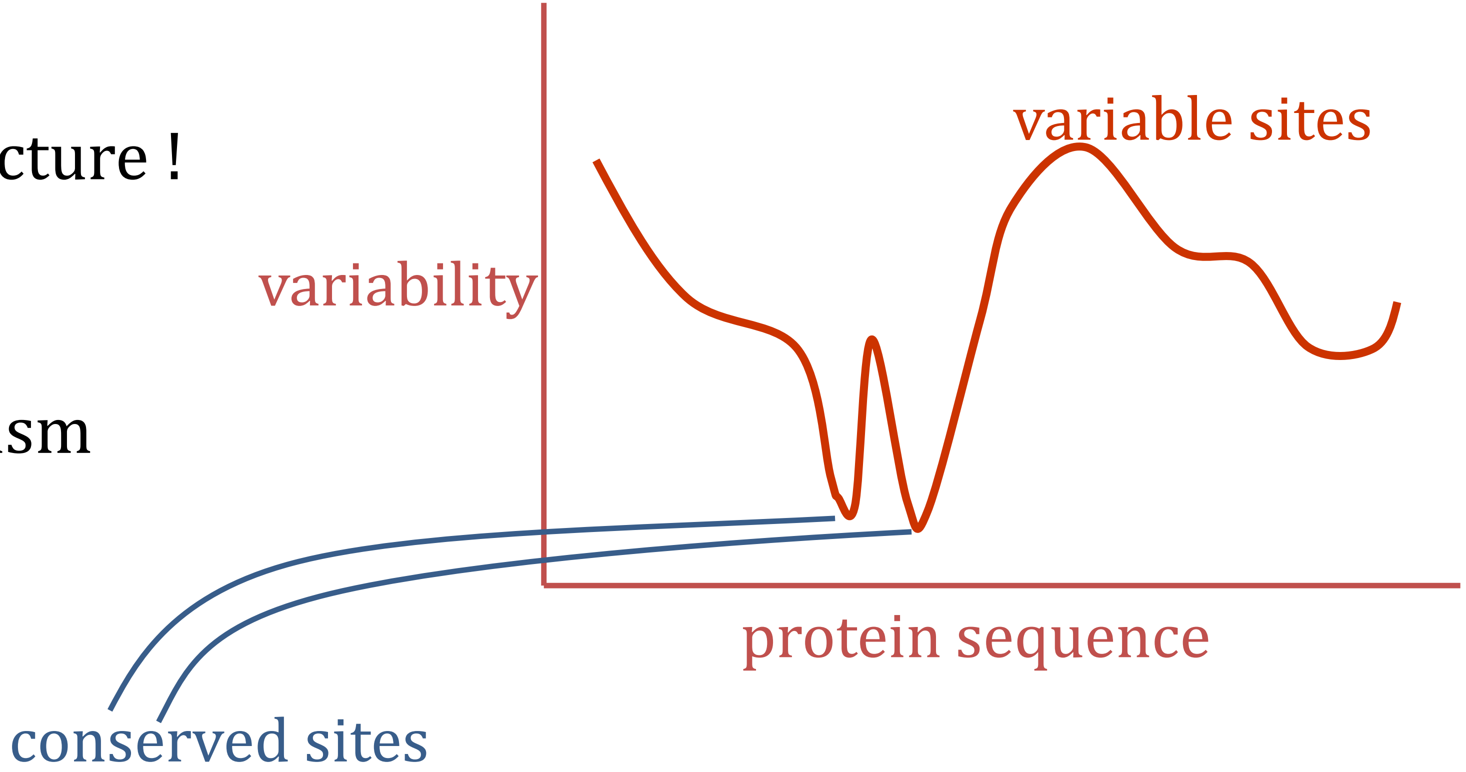
- see how conserved / variable they are

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

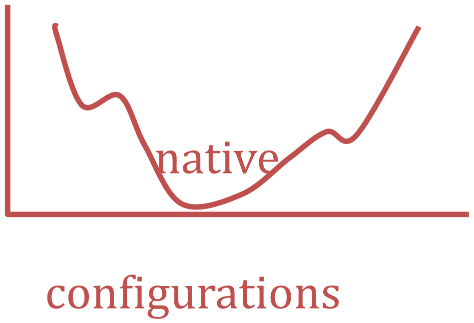
- function – protein must work
- stability – must be folded under normal conditions

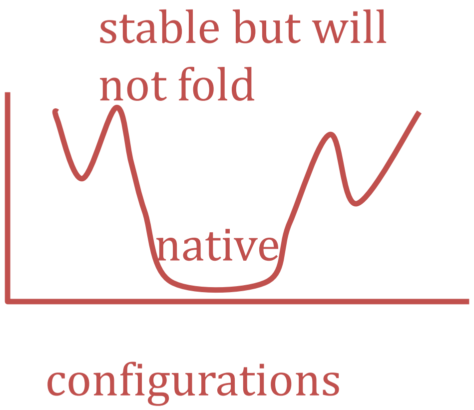
Less obvious, but simple

- kinetics – 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

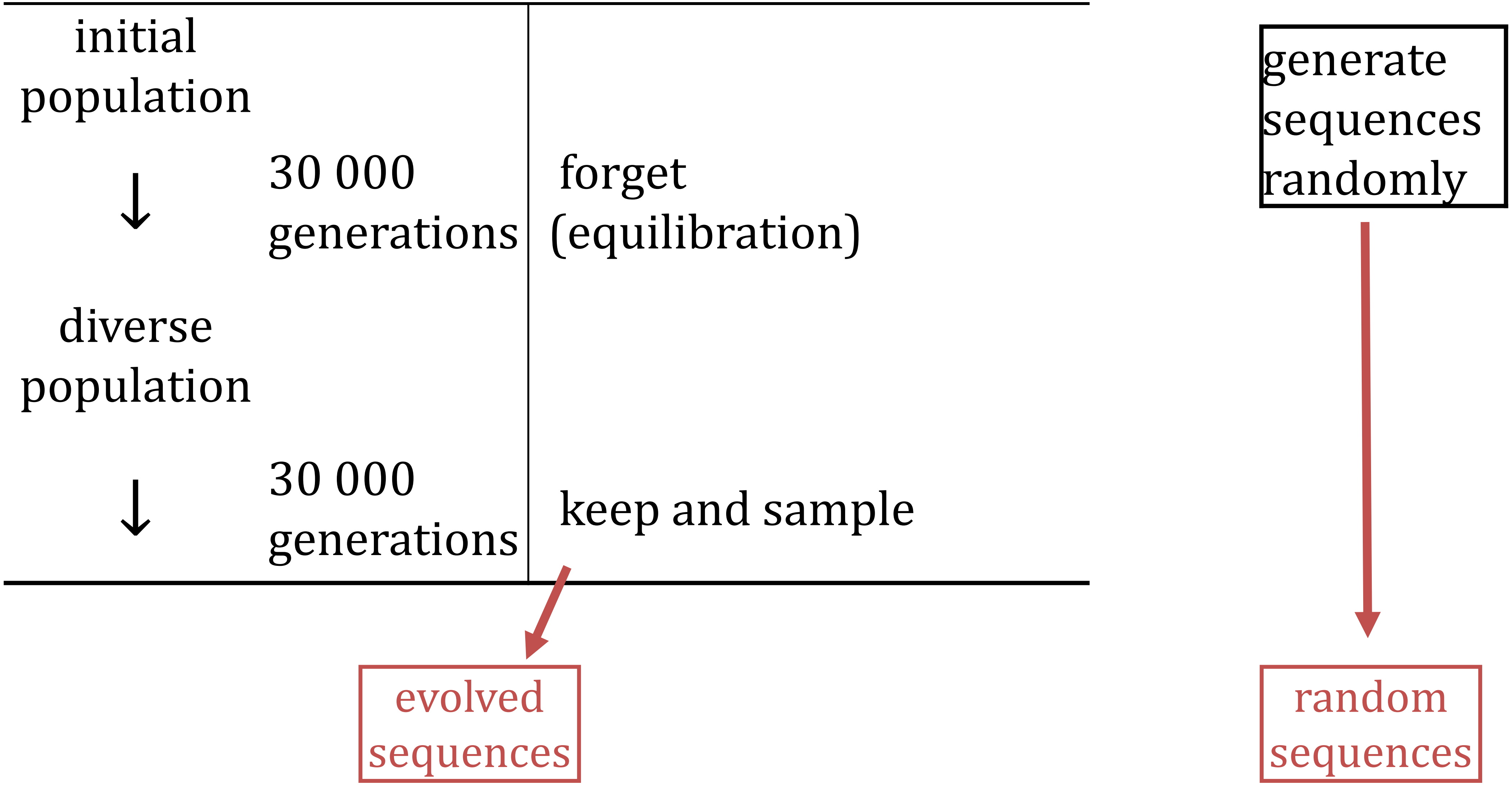
Normal procedure

- 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



initial population

↓ 30 000 generations | forget (equilibration)

diverse population

↓ 30 000 generations | keep and sample

generate sequences randomly

evolved sequences

random sequences

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 ?

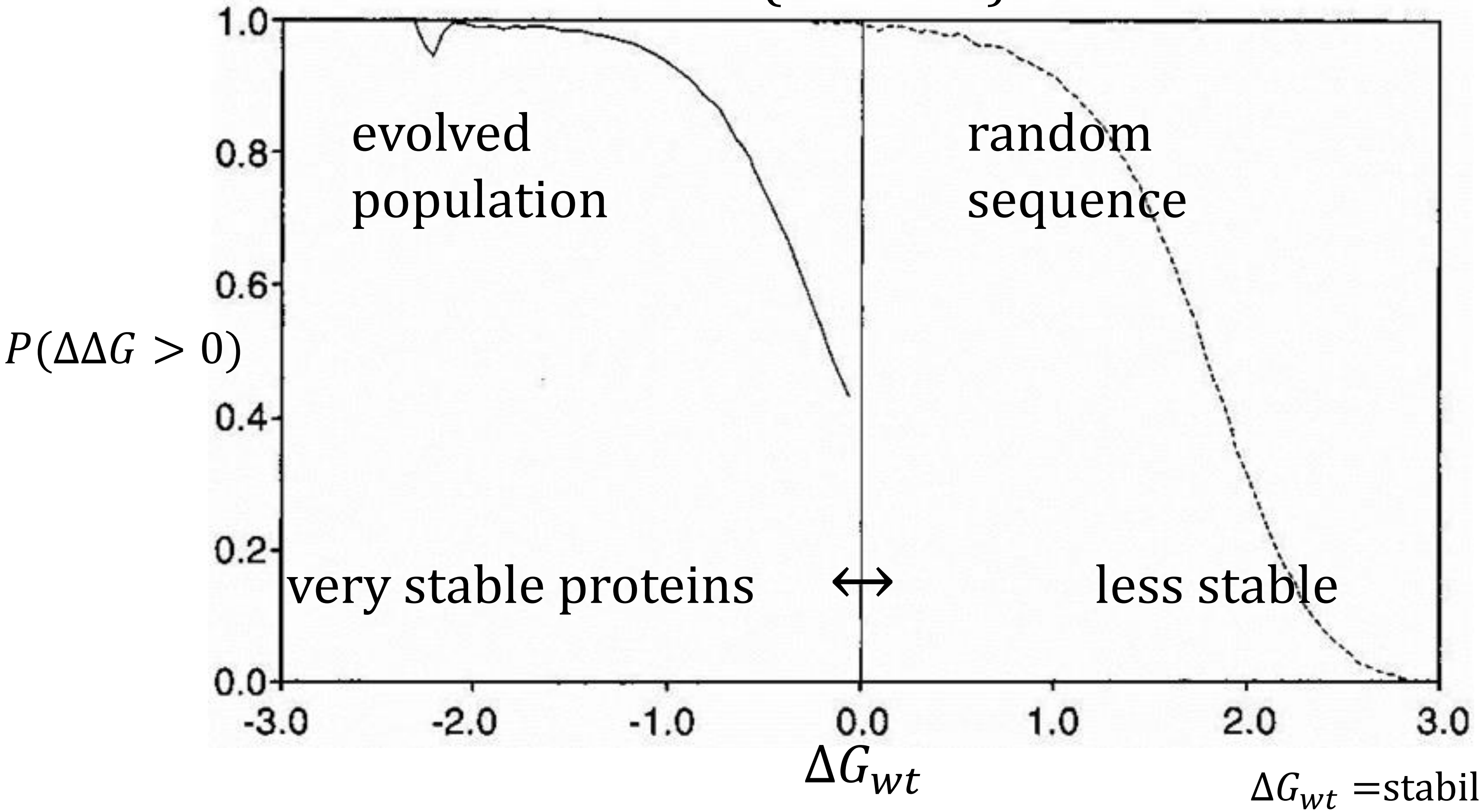
Evolved sequences must be more stable than random ones (obvious)

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

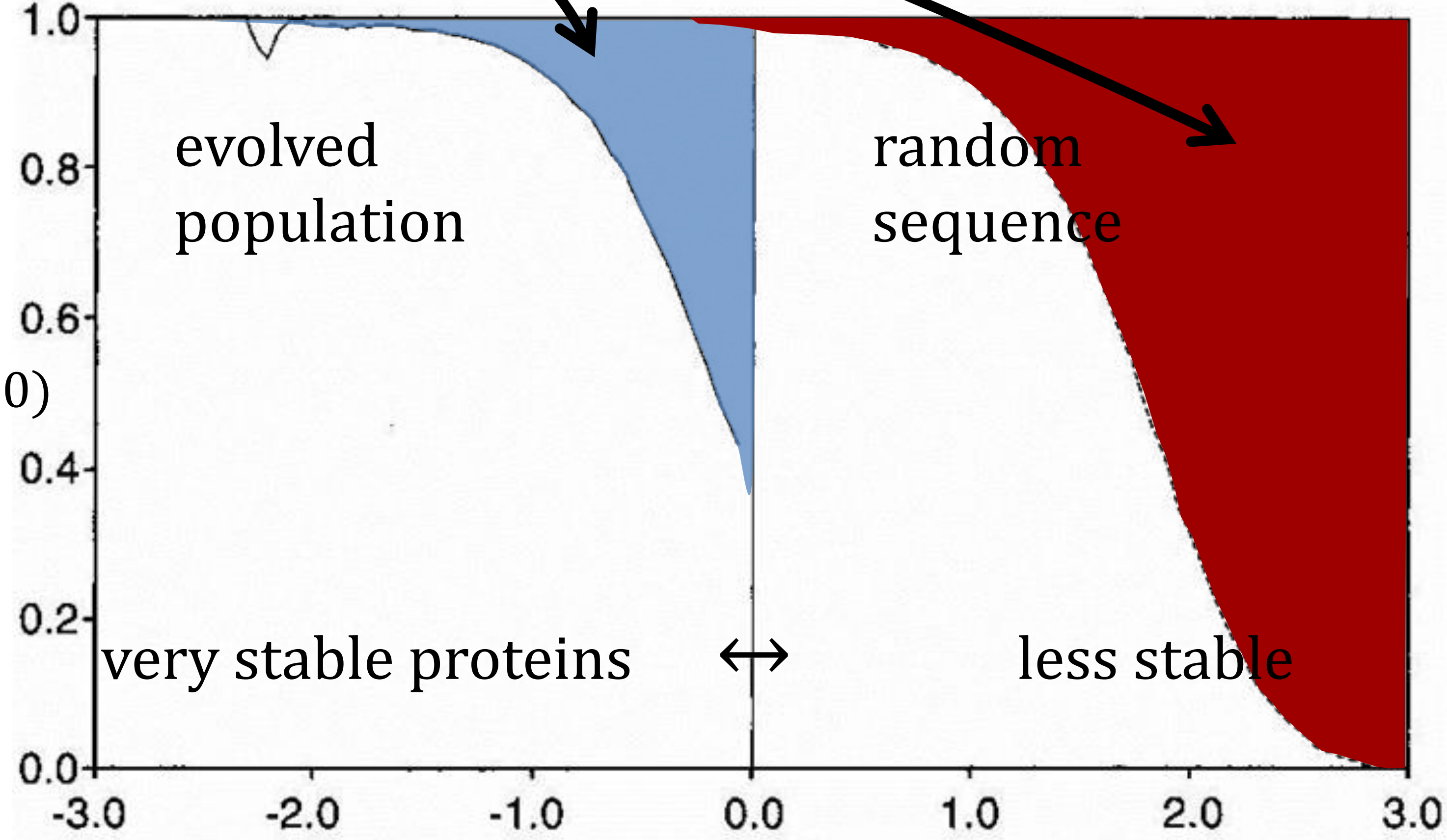
Taverna, DM and Goldstein RA, J. Mol. Biol. 315, 479-484 (2002)

Simulation results

Becomes more stable

probability of becoming less stable

$P(\Delta\Delta G > 0)$



evolved population

random sequence

very stable proteins ↔ less stable

ΔG_{wt}

ΔG_{wt} =stability before mutation

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

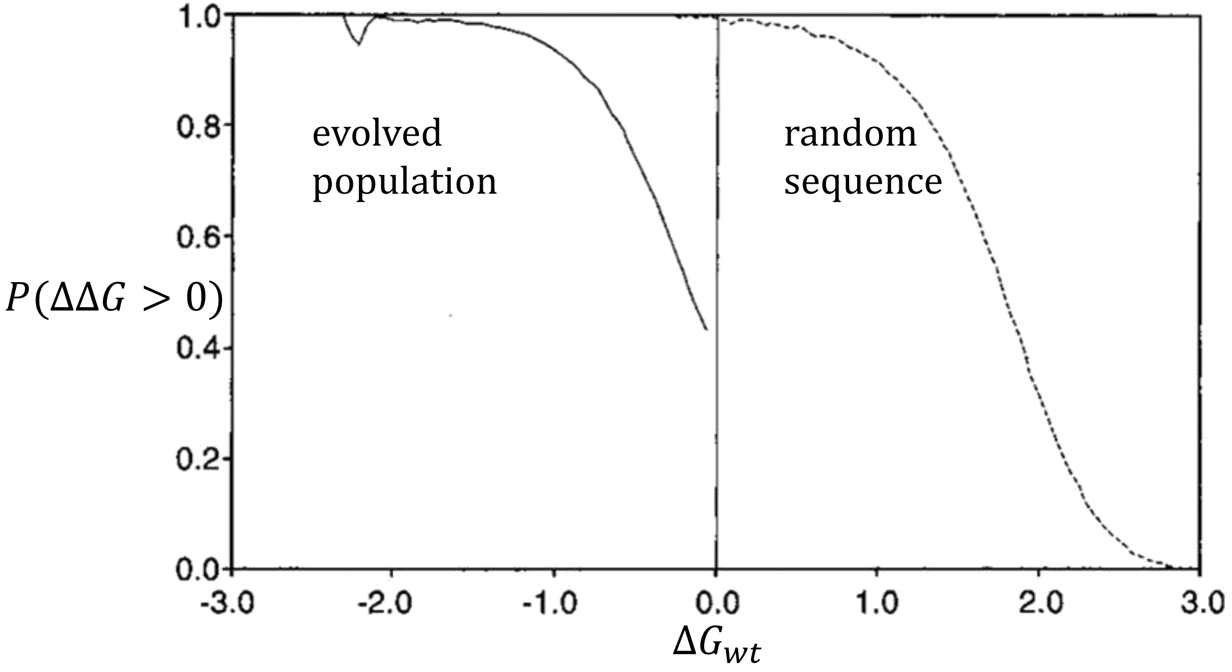
Interpreting results

random sequence

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

evolved sequence

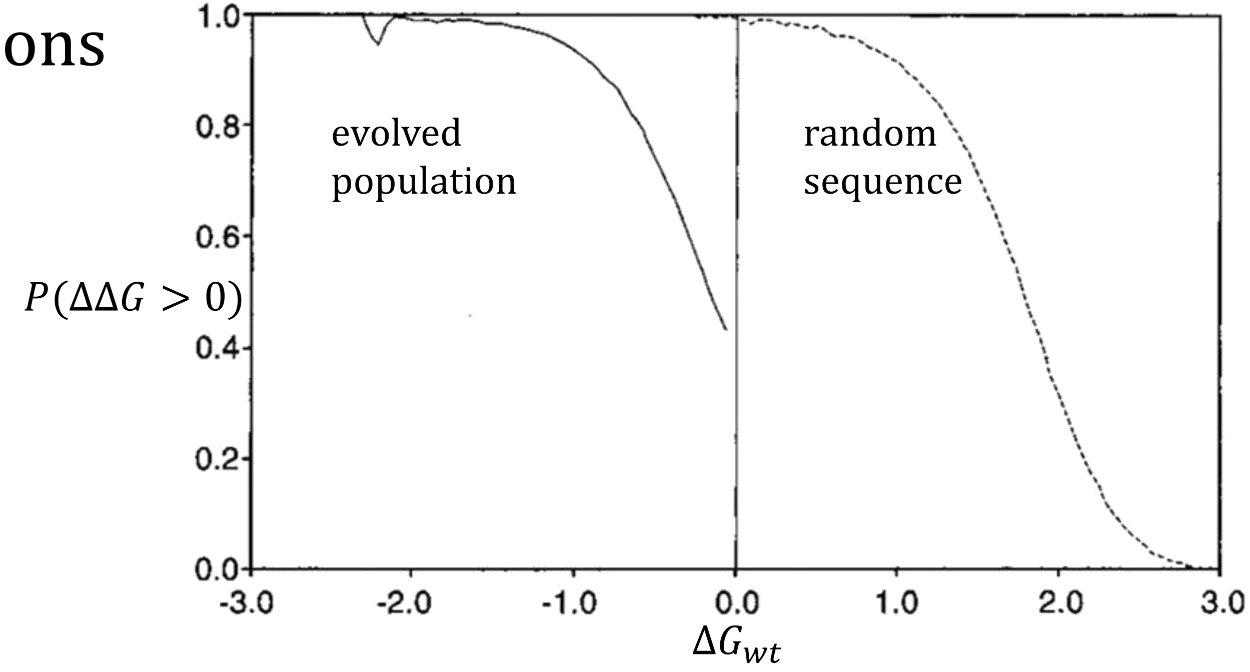
- very stable ?
  - cannot make better
- marginally stable ?
  - mutations often OK



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

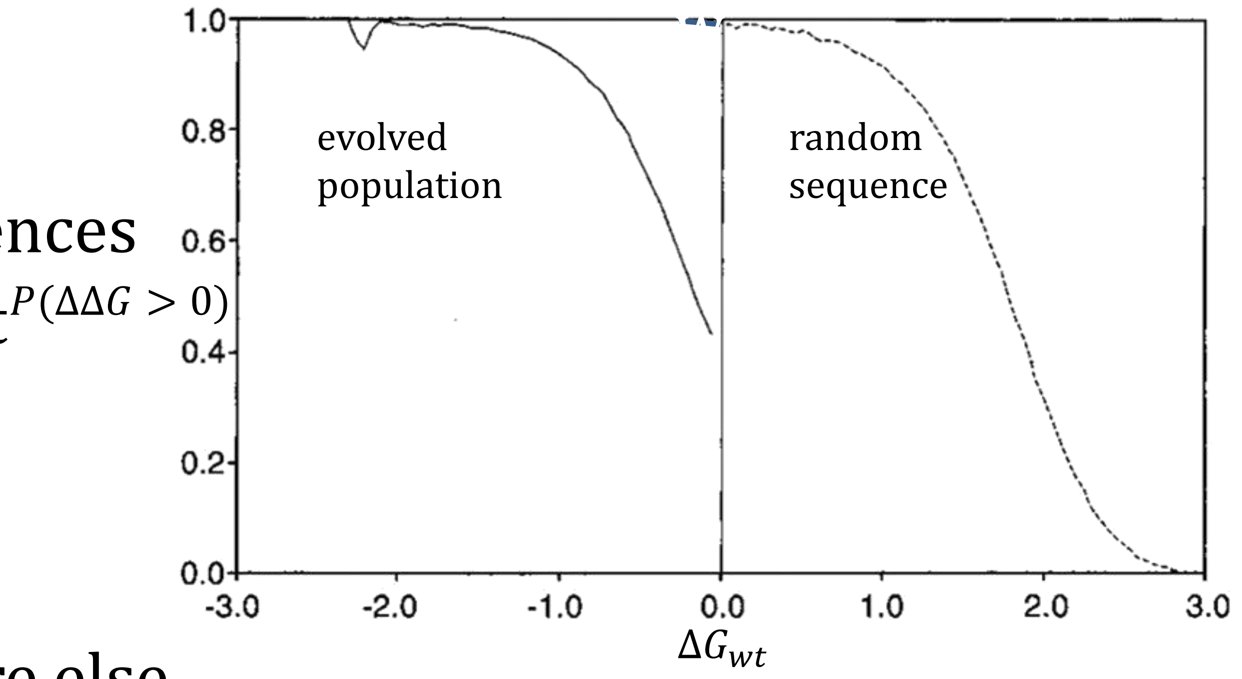
- moves population to left
- it should not change the fraction of stabilising mutations

selected

Simulation selected for stable sequences

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

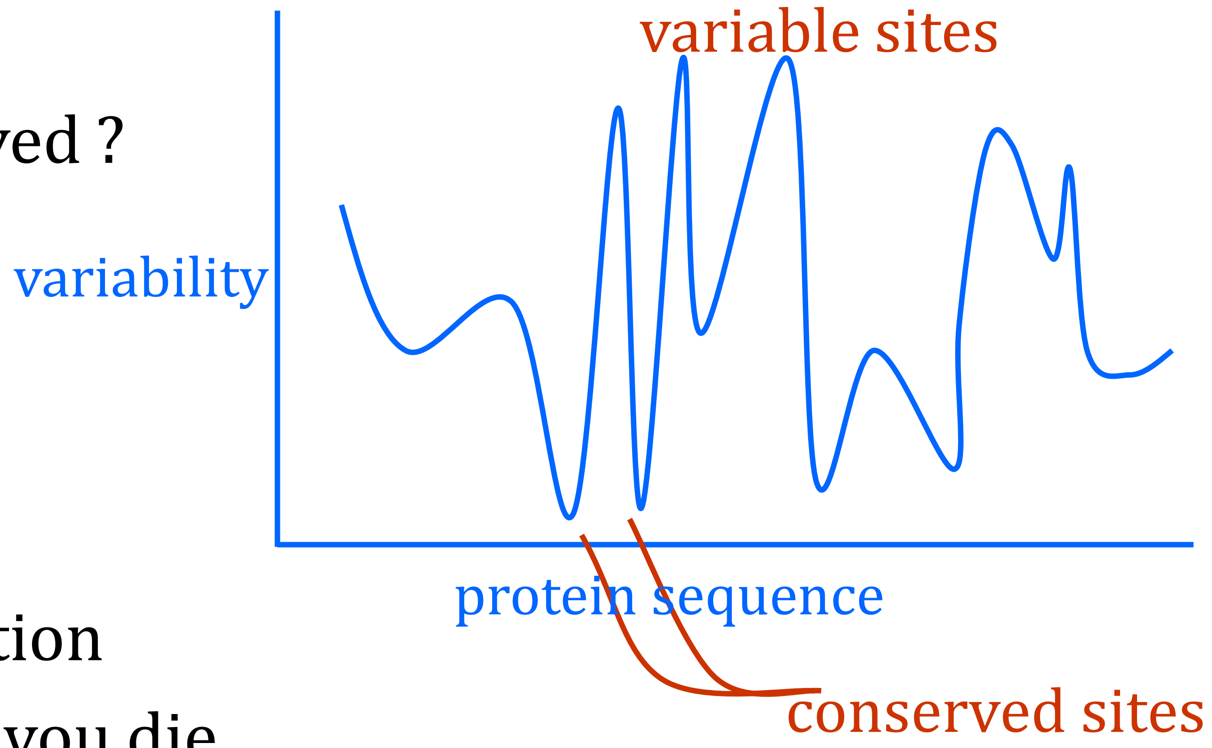
- effect must come from somewhere else



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

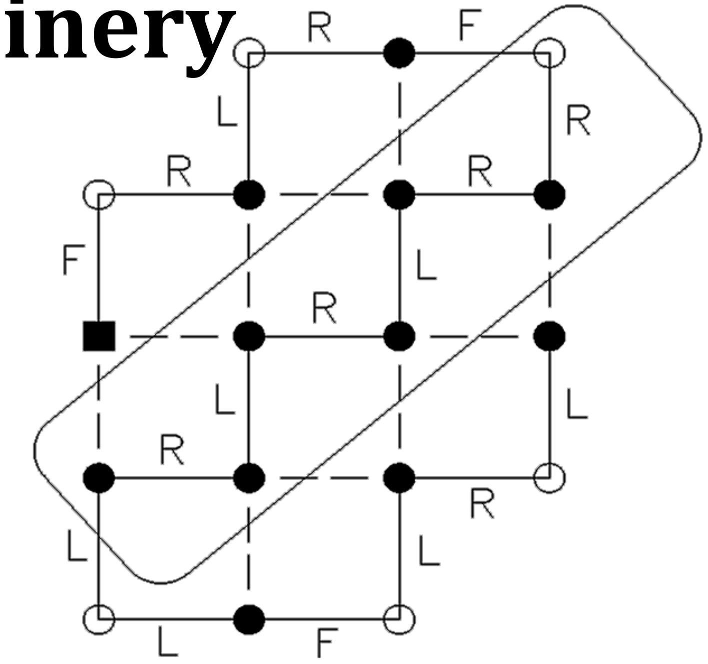
Basic idea

- 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

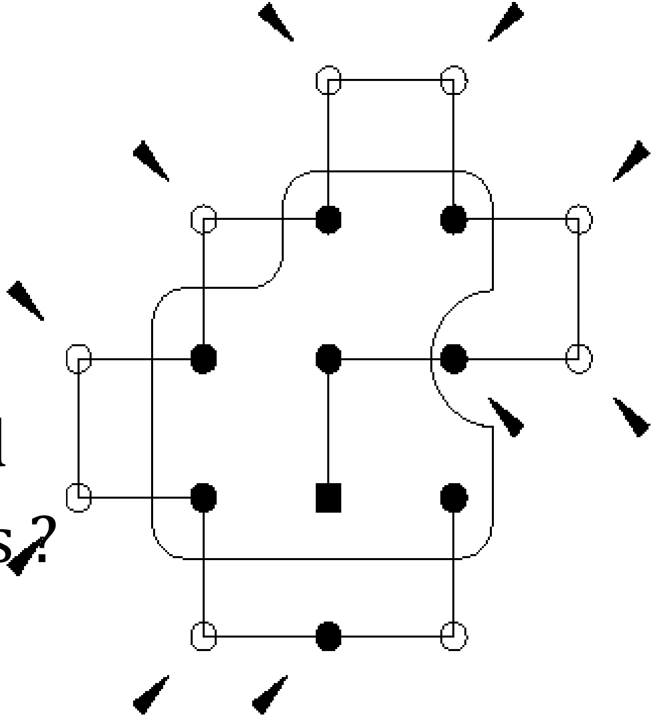
- many possible sequences

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



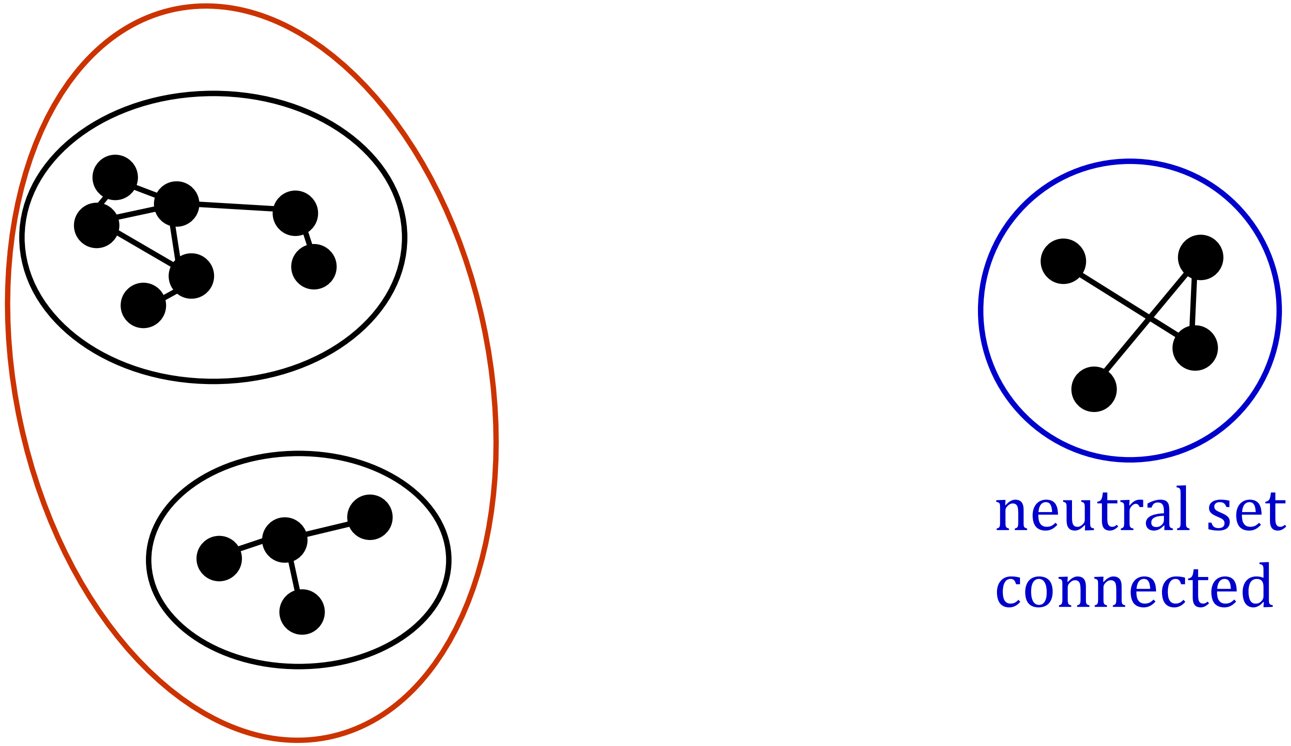
✓ sites where neutral mutations were found

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

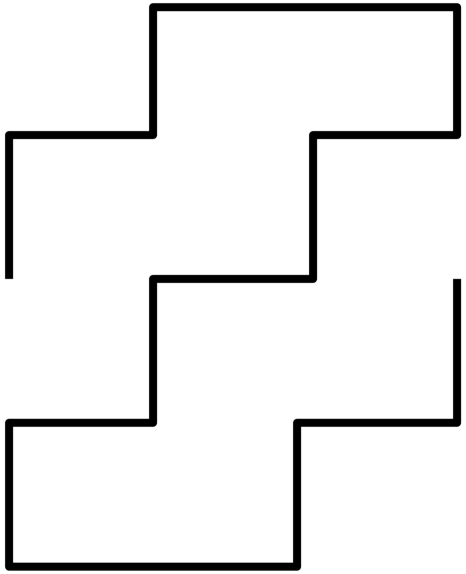


neutral set with two connected sets

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 ?

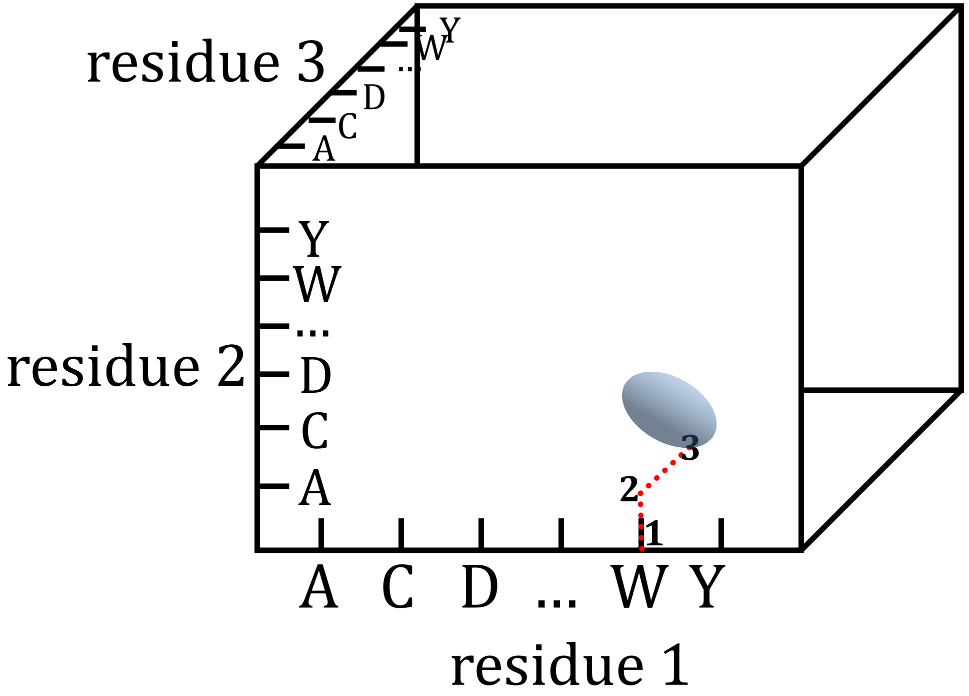
Mutation resistance revisited

Earlier slides

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

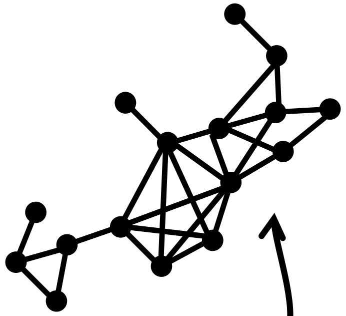
Alternative

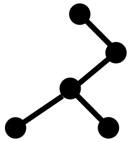
- think of sequence space
- a group of related sequences are a cluster in this space



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

Optimality

Spirit of Kimura (neutral evolution)

- most mutations are bad (pech gehabt)
- some mutations are rather neutral
  - will it become part of the genetic pool ? (fixation)

- Small population ? Maybe
- Big population ? Less likely

  $\frac{1}{2N_e}$ for some effective population N_e
- What if the mutation is a tiny bit harmful ? costs s
  - no problem
- Result ? Lots of small, slightly deleterious mutations - OK

Background of neutral evolution

DNA level (obvious)

- 64 codons / 20 amino acids / much redundancy
  - CUG / CUC both ile (+ many more)
- lots of mutations have no (not much) effect

Protein

- bit less clear
- we can change amino acids and
  - preserve structure
  - often function

Net effect

- we can make many mutations
- some do not affect the protein
- some protein effects are very small

Neutral evolution

Classical view (selective adaptation) explains life
- we are always trying to adapt to each other, environment ...
- there is some diversity when there is no cost (blue / brown eyes)

Alternative
- most mutations have no effect (neutral)
- if they far outnumber the selected mutations, they will dominate

Macroscopic
- brown eyes versus blue – not so surprising
- microscopic / molecular ?

Neutral evolution
- consequences ?
- predictions ?
- predictions at molecular level / simulations

Stability / Folding

- I must be stable at room temperature
- proteins in us must evolve to be stable under different conditions (organelles)
- extreme examples – bacteria
  - thermophiles, acidophiles, halophiles, …

- proteins are not really very stable (20 – 100 kJ mol^{-1}) – will come back

Function ? Obvious

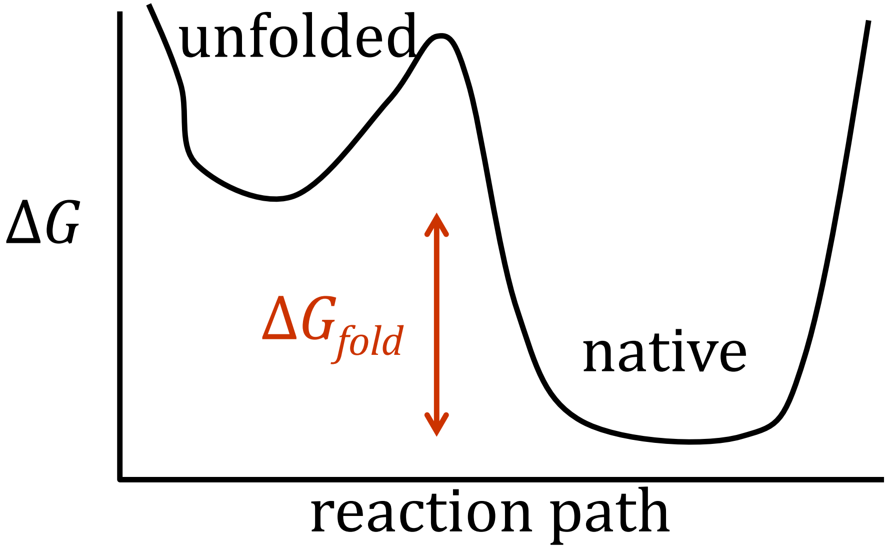
- If it is broken, you die

Protein stability

more work from same group*

Most proteins are NOT very stable (5 – 10 kcal mol^{-1})

- claims:
  - less stable, more flexible
  - easier to have chemical function



---

*Taverna, DM, Goldstein, RA, 2002, Proteins, 46, 105-109, Why are proteins marginally stable ?

Another model calculation

- 5×5 lattice 1081 conformations
- 20 amino acid types
- cannot visit all sequences, can visit all structures
- use a definition of foldable

$$\Delta G_{folding} = E_f + kT \ln\left(Z - \exp\left(\frac{-E_f}{kT}\right)\right)$$

3 simulations

1. long walk of one sequence
2. population
3. random sequences

Sidetrack for arguments

Goldstein's formula

- p_f probability of folded state $p_f = \frac{\exp\left(\frac{-E_f}{kT}\right)}{Z}$
- p_u probability of unfolded state
  - probability all states (1)− probability of folded $p_u = \frac{\sum_i \exp\left(\frac{-E_i}{kT}\right) - \exp\left(\frac{-E_f}{kT}\right)}{Z}$

$$\frac{p_f}{p_u} = \frac{\exp\left(\frac{-E_f}{kT}\right)}{\sum_i \exp\left(\frac{-E_i}{kT}\right) - \exp\left(\frac{-E_f}{kT}\right)}$$

$$= \frac{\exp\left(\frac{-E_f}{kT}\right)}{Z - \exp\left(\frac{-E_f}{kT}\right)}$$

Getting free energy expression

$$\Delta G = -kT \ln\left(\frac{p_f}{p_u}\right)$$

$$= -kT \ln\left(\frac{\exp\left(\frac{-E_f}{kT}\right)}{Z - \exp\left(\frac{-E_f}{kT}\right)}\right)$$

$$= -kT \ln \exp\left(\frac{-E_f}{kT}\right) + kT \ln\left(Z - \exp\left(\frac{-E_f}{kT}\right)\right)$$

$$= E_f + kT \ln\left(Z - \exp\left(\frac{-E_f}{kT}\right)\right)$$

folded $\overset{\Delta G}{\rightleftharpoons}$ unfolded

and we usually write

$$\Delta G = -RT \ln \frac{[\text{folded}]}{[\text{unfolded}]}$$

Simulation (long walk)

Take viable sequence

- mutate
    - if (foldable)
        - keep
    - else
        - retain old sequence

Simulation (population)

- Take 3000 identical sequences
- mutate

- calculate $\Delta G_{folding}$ for all members
- kill (remove) non-folders
- copy random survivors to keep population at 3000

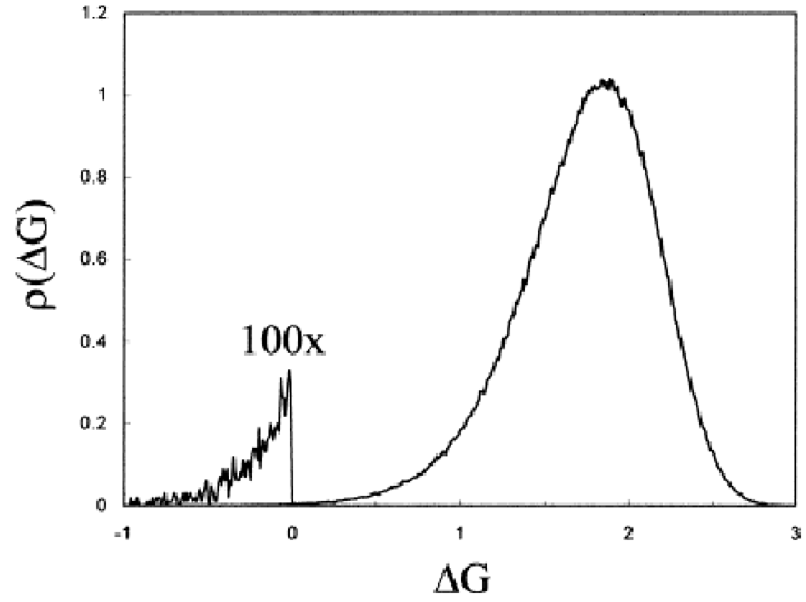
Stability of results

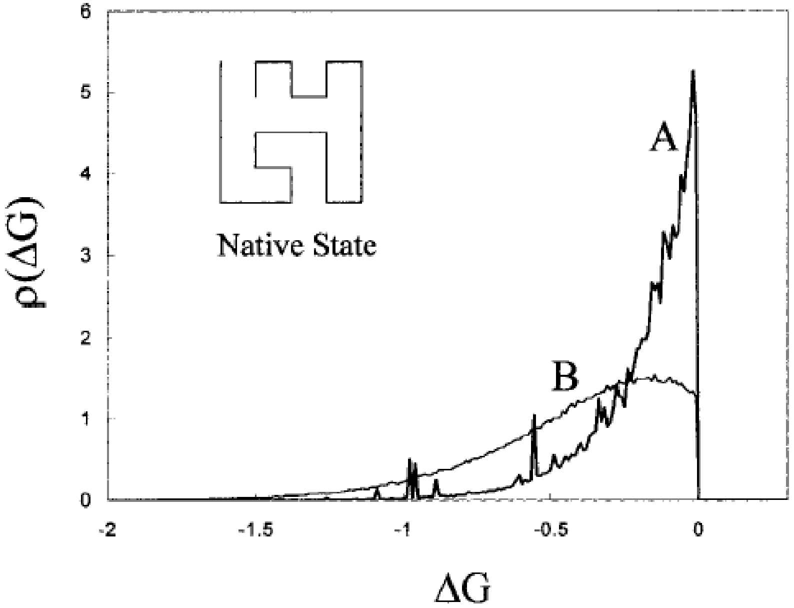
What is the result

- from random sequences ? (left)
- from a long walk (right A)
- from a population (right B)

Sequences become more stable

- but barely so





Taverna, DM, Goldstein, RA, 2002, Proteins, 46, 105-109, Why are proteins marginally stable ?

Where does the population result come from ?

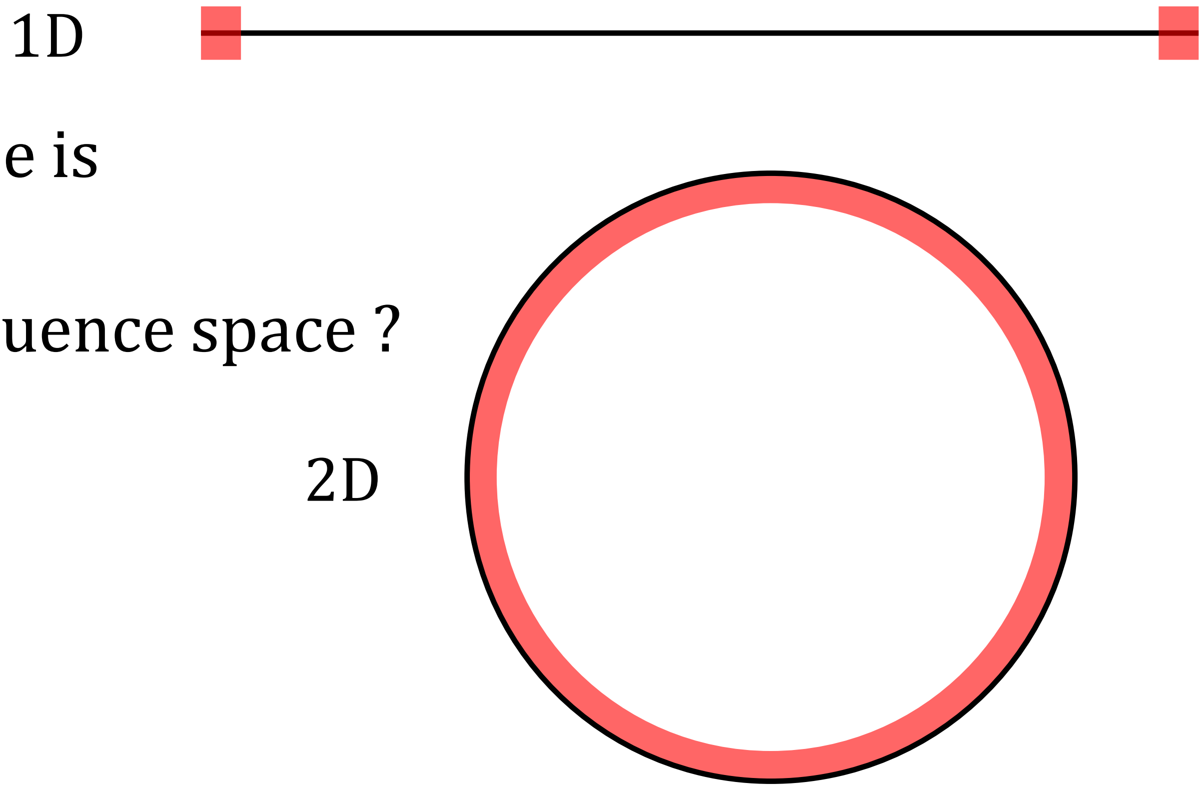
Proteins die if they are unstable

- the population moves to folding sequences (this is selected)
- there is no force to make them more stable

- high dimensional object arguments / population phenomena
  - explain the population result

dimensions / surface area

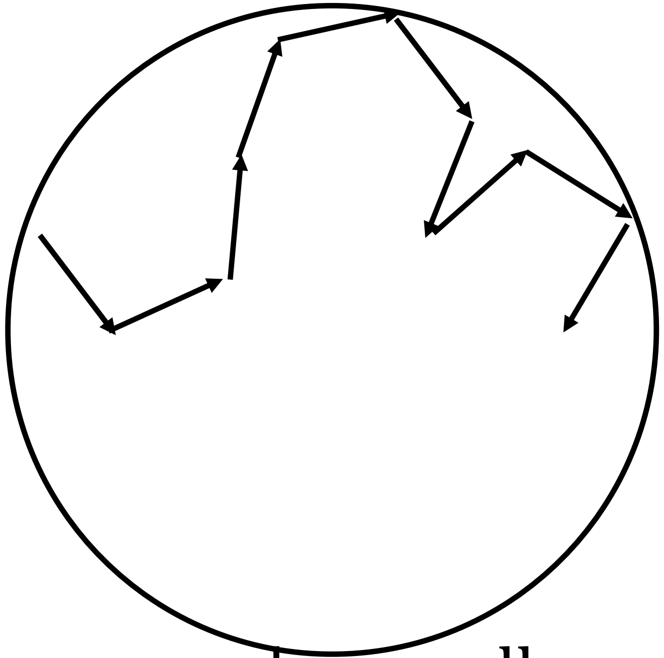
- 3D more near surface
- ...
- high-D most of volume is near surface
- dimensionality of sequence space ?



Walk versus Population

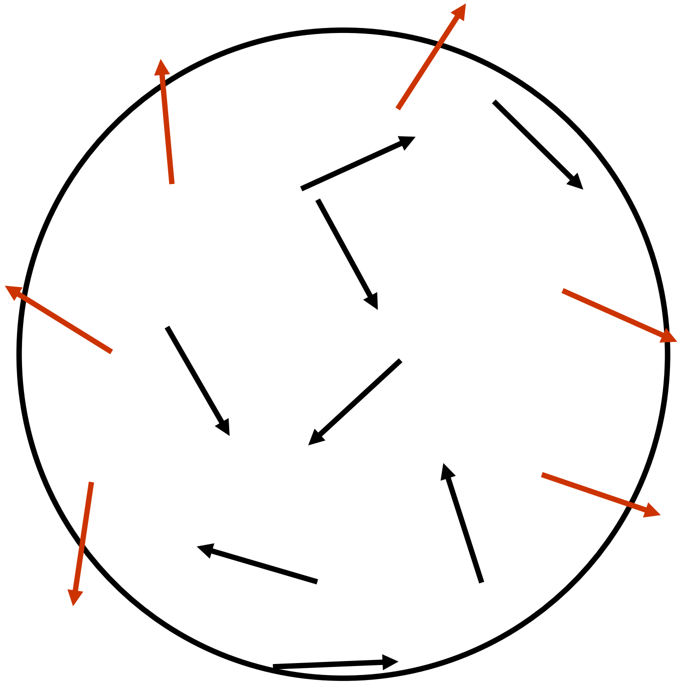
High dimensional objects

- high proportion near to surface



long walk

- sequences bounce around near surface



population

- sequences near surface removed, others reproduce

Population acts as if there is a sink removing most unstable proteins

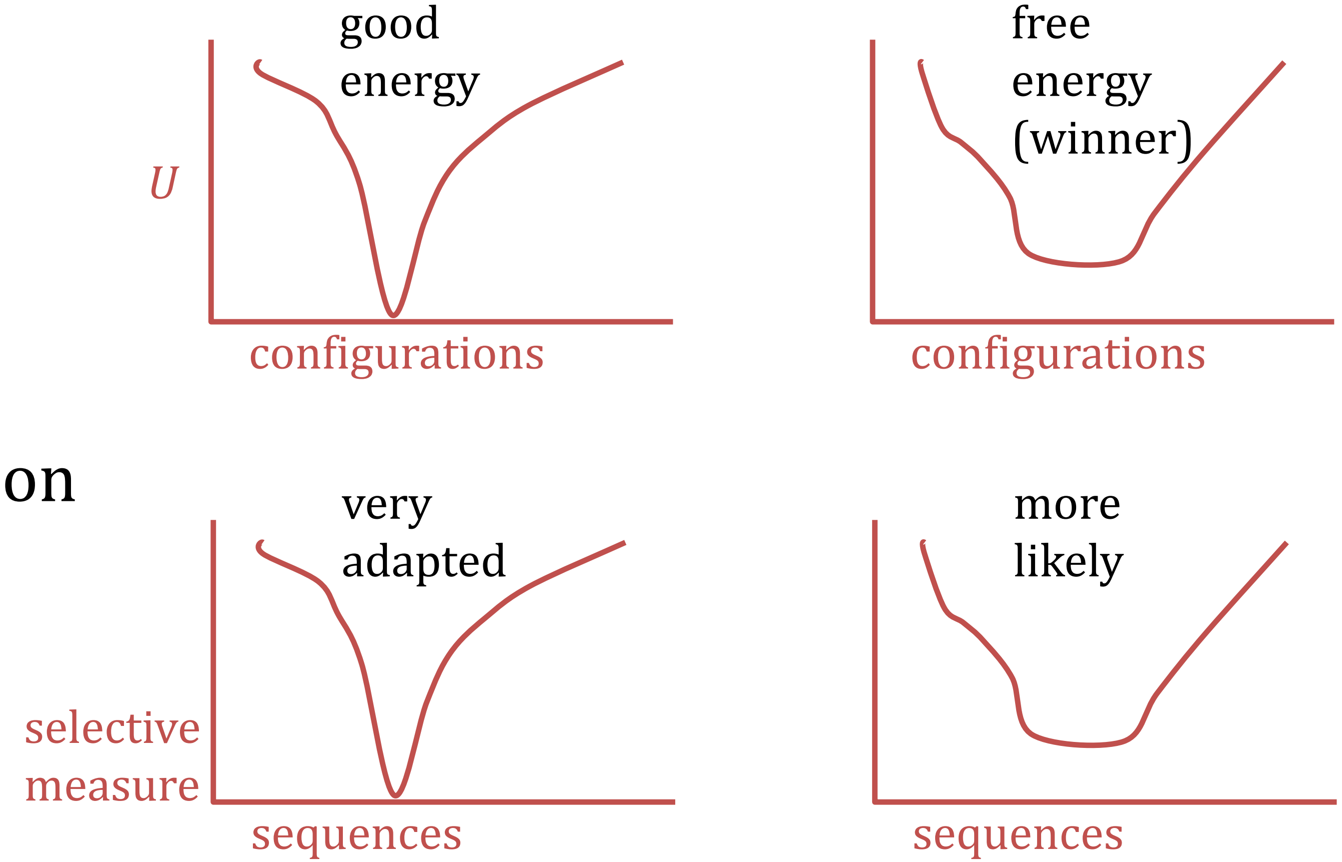
Results give marginally stable proteins

- no mention of function
- arguments purely statistical

Analogy: evolution and free energy

energy /free energy minima

evolutionary version



- evolution is adaptive, but subject to statistical effects
- statistical effects may look like evolutionary pressures (mutation resistance, stability)

Summary

First lattice lectures

- one can do Monte Carlo simulations

Now

- there are other types of simulation

Trying to interpret world in terms of evolutionary pressure not always justified

Evolutionary implications

- something looks Darwinian really reflects structure of sequence / structure space