The Boltzmann factor

Many molecules are flexible to some extent, and a molecule can be expected to achieve a variety of different conformations with different associated energies.

Intuitively we would expect the conformations, or states, with low energy to be observed more frequently than those with higher energy.

The Boltzmann factor gives us a way to quantify the likelyhood that a particular conformation will be observed, given the energy of that conformation.

The Boltzmann factor

If a multi-state system is in thermodynamic equilibrium at temperature T, the probability of state i is proportional to

$$\frac{1}{e^{E/kT}} \tag{1}$$

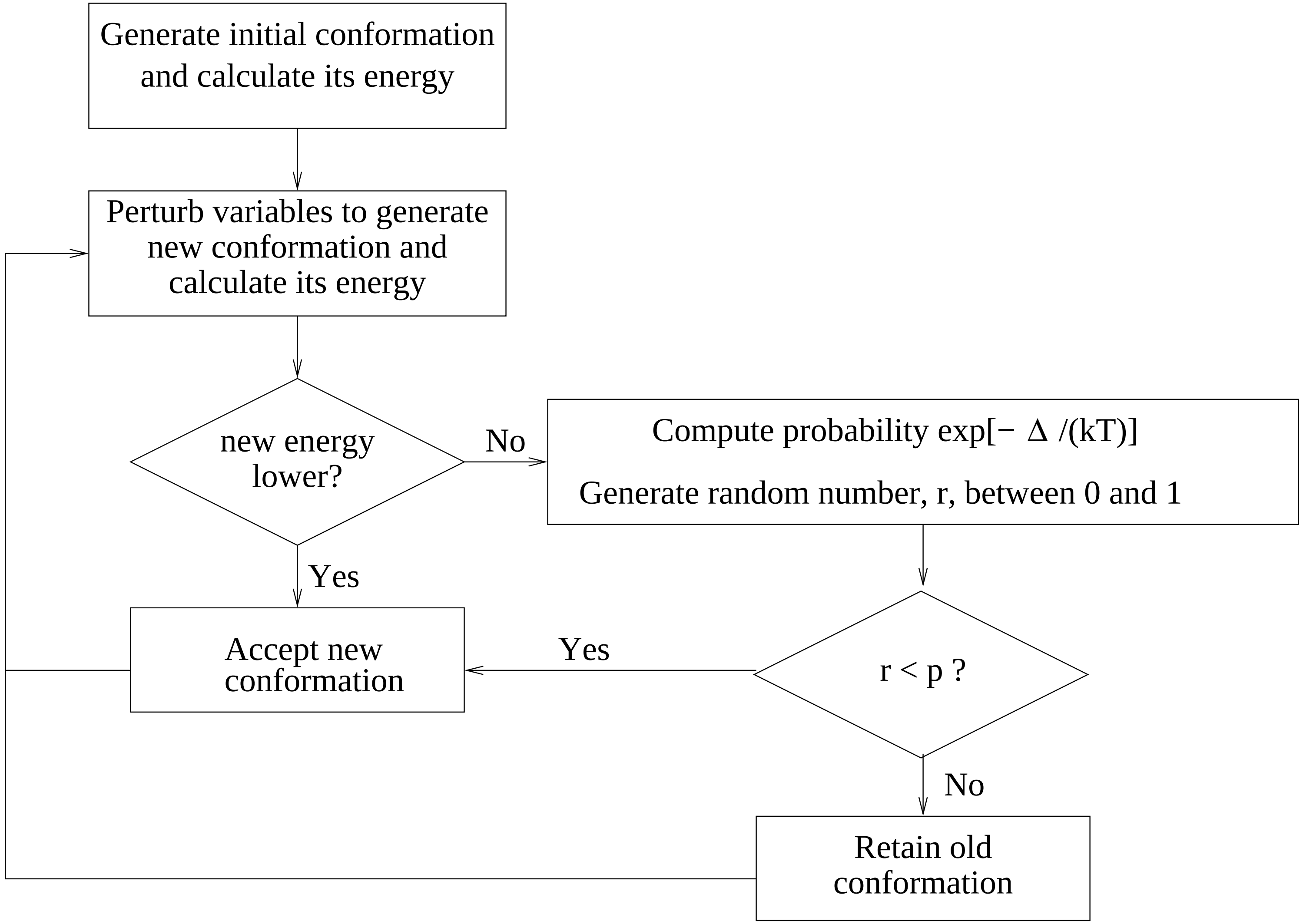
where E is the energy of state i and k is Boltzmann's constant.

This quantity, often written as $e^{(-E/kT)}$, is *the Boltzmann factor*.

As E increases, the Boltzmann factor decreases, meaning that it is progressively less likely that higher energy states will be attained.

As T increases, the Boltzmann factor increases, meaning that at higher temperatures it is more likely that a state with higher energy will be found.

Metropolis procedure



Generate initial conformation and calculate its energy

Perturb variables to generate new conformation and calculate its energy

new energy lower?

Compute probability $exp[- \Delta /(kT)]$

Generate random number, r, between 0 and 1

$r < p$?

Accept new conformation

Retain old conformation

Why build model structures?

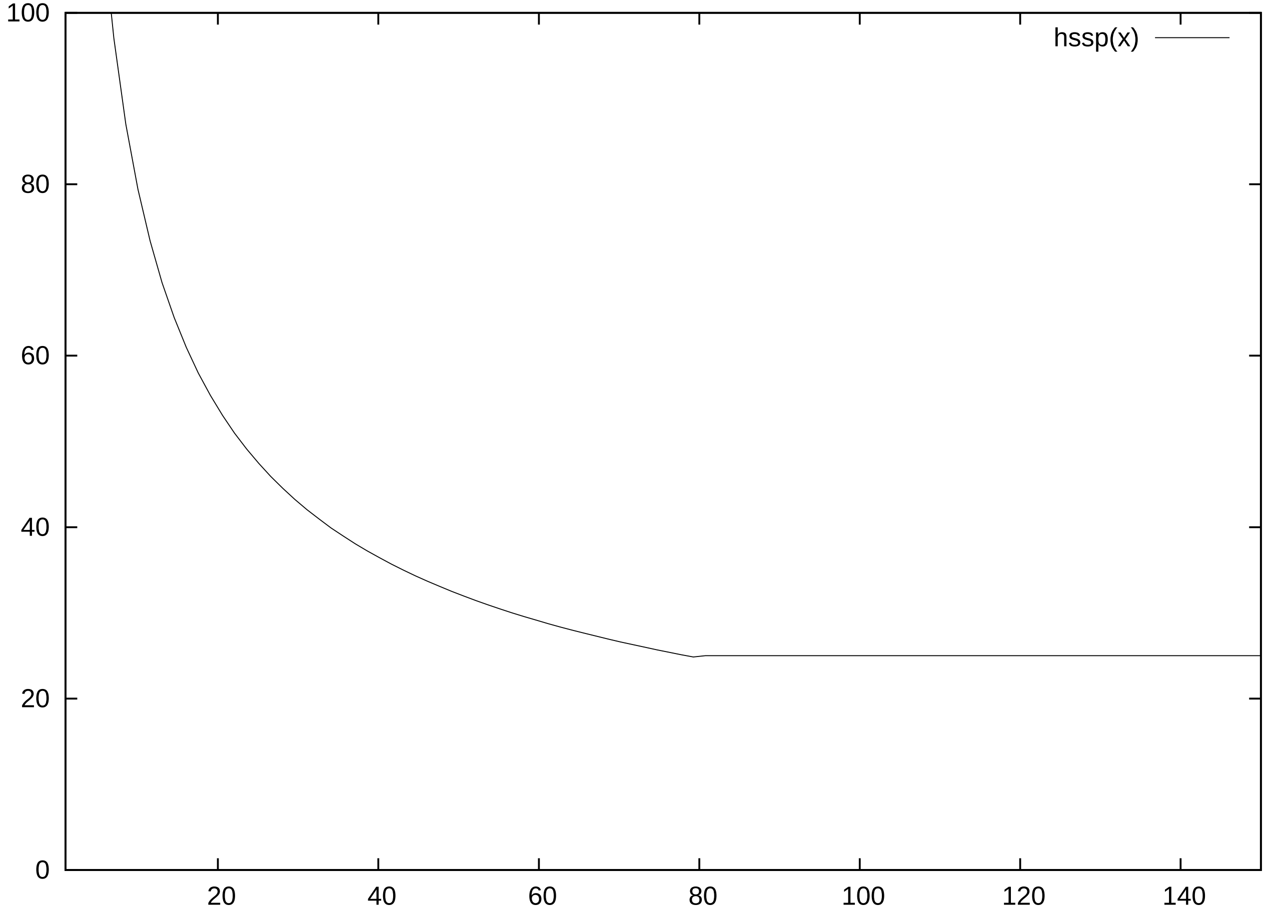
Knowledge of a protein’s three-dimensional structure is vital to a full understanding of the molecular basis for its biological function.

We want to understand the function of all proteins encoded by a genome, therefore we would like to know all of their 3-D structures.

Experimental techniques for determining protein structure are relatively slow and expensive, so we look to modelling as a way of extending the set of 3-D structures.

Modelling can also be used in protein engineering when designing proteins for therapeutic applications.

HSSP-curve



[Sander C. and Schneider, R., Proteins: Structure, Function and Genetics, 1991, 9:55-68]

“HSSP-curve”

- Shows the length-dependent threshold for significant sequence identity.
- Proposed by Sander and Schneider (1991) and revised by Rost (1999).
- Above the curve, identifing true positives is easy.
- Just below the curve, the number of false positives rises rapidly; distinguishing between true and false positives in the “twilight zone” is difficult.

(HSSP stands for “Homology-derived Secondary Structure of Proteins”)

Comparative modelling strategy

- identify a known structure that is predicted to be similar;
- align sequences;
- predict structurally conserved regions, and locations of insertions and deletions (sometimes called “indels”);
- build model backbone structure
  - — copy predicted conserved main chain regions from template structure,
  - — remodel loops with insertions or deletions;
- add side chains to the modelled main chain;
- evaluate and refine model.

Using known substructures in protein crystallography

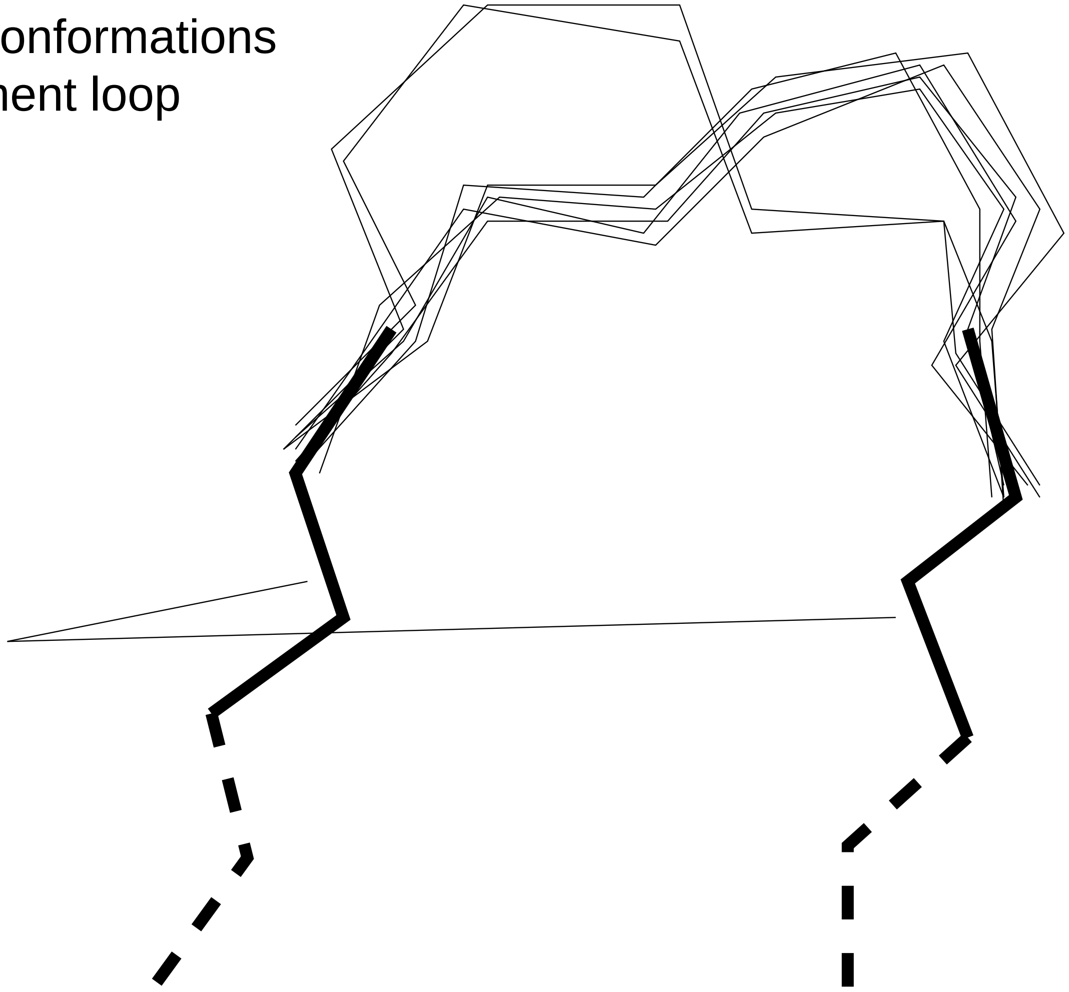
Jones, T.A. and Thirup, S. (1986)
The EMBO Journal, vol. 5, pp 819-822.

Electron density map interpretation is made easier by fitting regular α-helices and strands into the map.

This building-block approach to protein modelling can be extended to include **all** main chain fragments.

For example, a model of retinol binding protein was built using fragments from only three other proteins. A model with Cα atoms matching within an R.M.S. error of 1 Å was built using only 15 fragments.

Fragment-fitting: an approach to remodelling loops



Candidate conformations
for replacement loop

Structurally
conserved
framework

Side chain rotamers

There is an extremely large number of possible combinations of side chain conformations — infinite if we consider side-chain bonds to be continuously variable.

For practical purposes the search space can be discretised by considering a finite set of possible torsion angles for each side-chain.

The distribution of side chain conformations falls into statistically significant clusters. By using representative side chain conformations, or **rotamers**, the vast combinatorial search space can be greatly reduced.

> Ponder, J.W. and Richards, F.M. (1987)
> J. Mol. Biol., vol. 193, pp 775-791.

Fold recognition

The idea behind “threading”:

Imagine a wire wound into the shape of a known protein’s main chain “fold”.

Imagine next that our new sequence is represented by beads that are “threaded”, in order, onto the wire, and are pushed along the wire.

At each step, a score is calculated based on which residues are adjacent in space, which residues are buried, etc.

Repeat this process for each different known fold.

A high score indicates that the sequence is compatible with that fold.

Approaches to fold recognition

Profiles

e.g. Bowie et al. (1991) Science, 253:164-170

Pairwise potentials

e.g. Sippl and Weitckus (1992) Proteins, 13:258-271
e.g. Jones et al. (1992) Nature, 358:86-89
e.g. Jones (1999) J. Mol. Biol., 287:797-815 — GenTHREADER

pairwise pseudo-energy terms

$$\Delta E_k^{ab} = RT \ln(1 + m_{ab}\sigma) - RT \ln\left(1 + m_{ab}\sigma \frac{f_k^{ab}(s)}{f_k(s)}\right)$$

solvation potentials

$$\Delta E_{solv}^{a}(r) = -RT \ln\left(\frac{f^{a}(r)}{f(r)}\right)$$

Threading the peptide through the groove

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

...GACPKYVKQNTLKLATGMRNVPEKQTRGLFGA...

Scoring function

- pairwise frequencies
- steric overlap and quality of fit
- hydrogen bonds
- positive and negative charges
- buried hydrophobic side chains
- exposed hydrophilic side chains