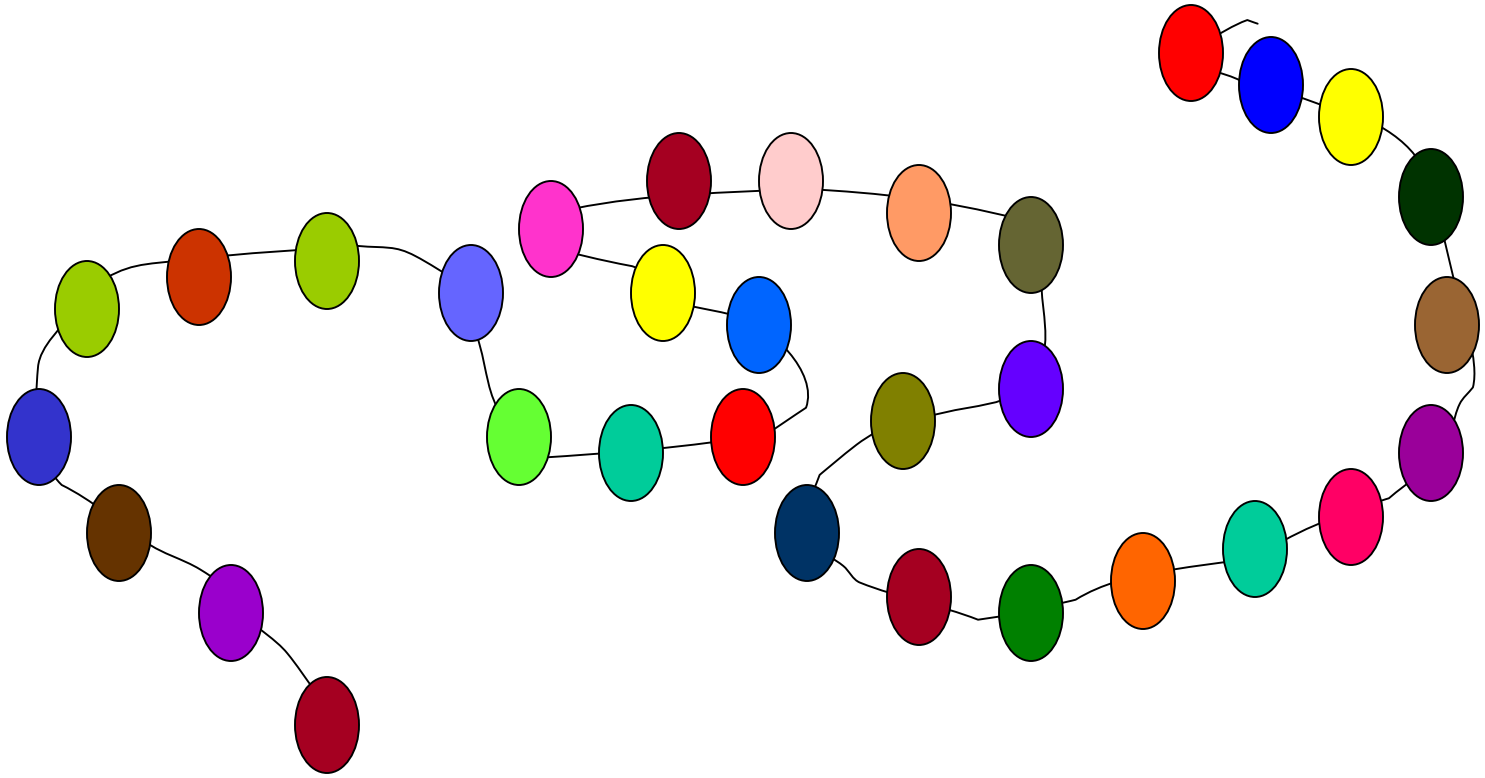
Protein folding

From linear information to a well defined three dimensional folds.

The existence of a unique fold is a BIG surprise

Proteins are heterogeneous polymers: many (20) types of monomers



Free energy driving force for protein folding

- “phase separation” to amino acids that “hate” water (hydrophobic) and amino acids that are well solvated in water (hydrophilic)
- Secondary structure is determined by short range hydrogen bonding (no or very little energetic gain compared to water solvation).

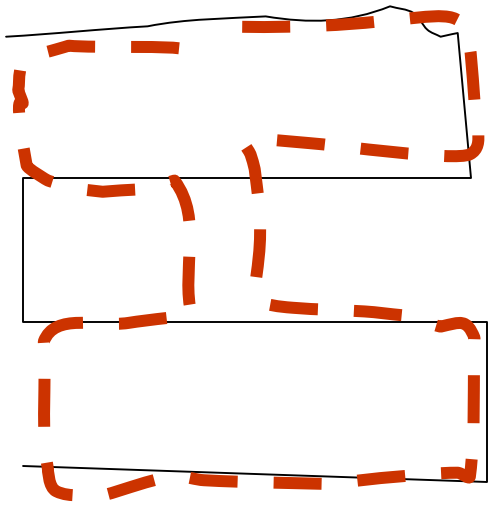
Proteins are one dimensional polymer that adapts well define three dimensional structure

- Heterogeneity is essential for a unique fold
  - Name another biological heterogeneous polymer with unique structure
  - Name a biological ("almost") homogeneous polymer without a specific unique fold
- Extreme views of amino acid heterogeneity that assists in understanding determinants of unique structure
  - Only two types of amino acid H (Hydrophobic and P (polar)
  - Infinite number of types (the random energy model)

A simple argument why heterogeneity is necessary to fold

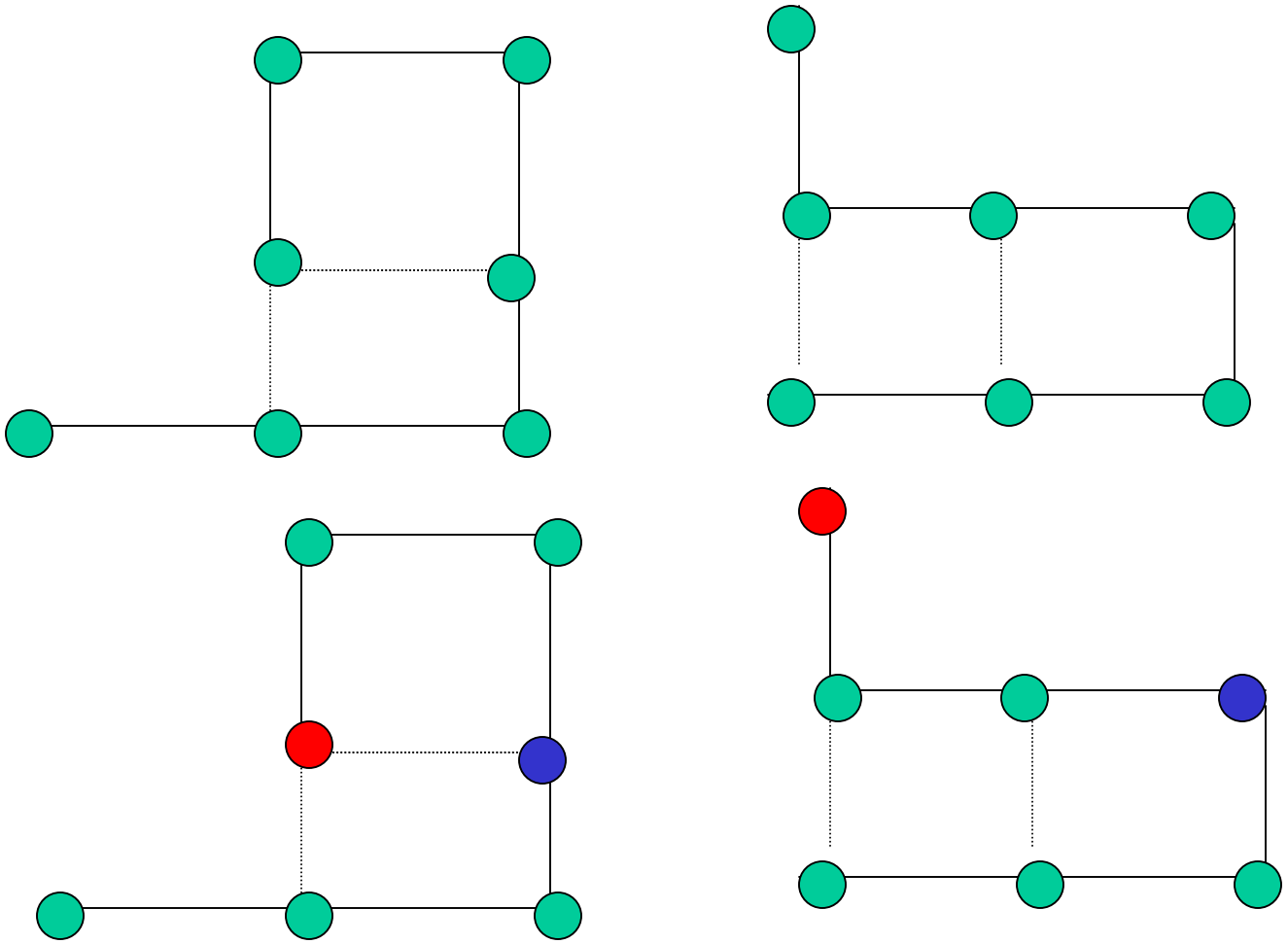
We "score" a structure or a shape by an energy function

$$\frac{1}{2}\sum_{i,j} h_{ij}\left(r_{ij}-r_c\right)u_{i,j}\left(\alpha,\beta\right)\equiv\sum_{c} u_c\left(\alpha,\beta\right) \qquad \forall i,j$$

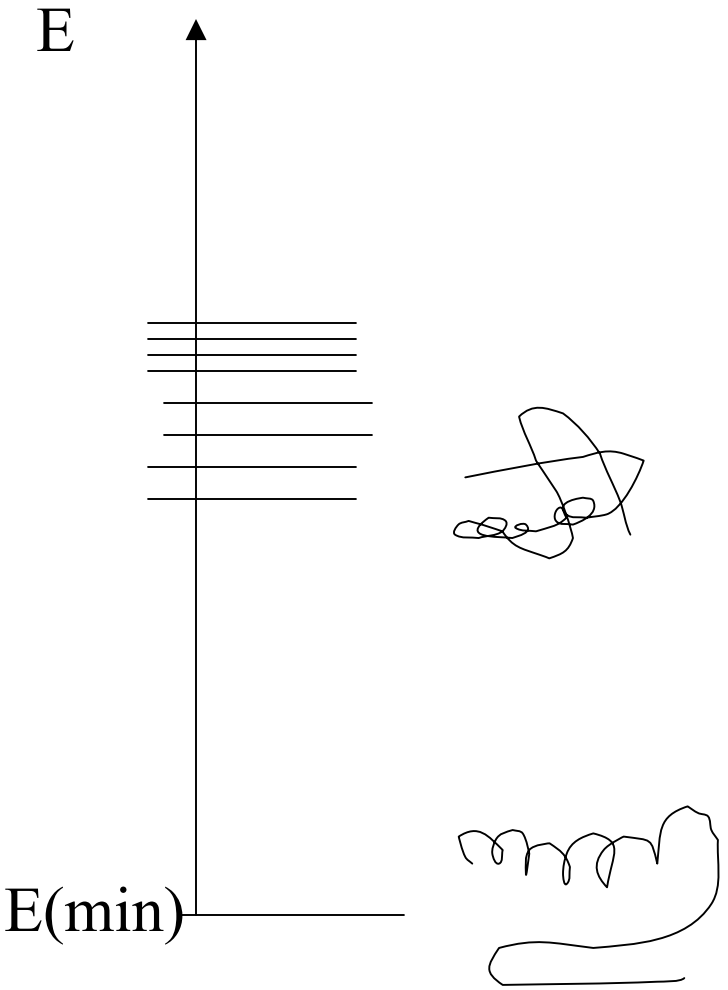


Both shapes (red & black) have roughly the same volume, the same number of contacts, and the same type of geometrical interactions. If all the monomers are the same the energy will be the same

To differentiate between the two structures below we must have more than one color



Need energy gap to create stable three dimensional shape



The random energy model

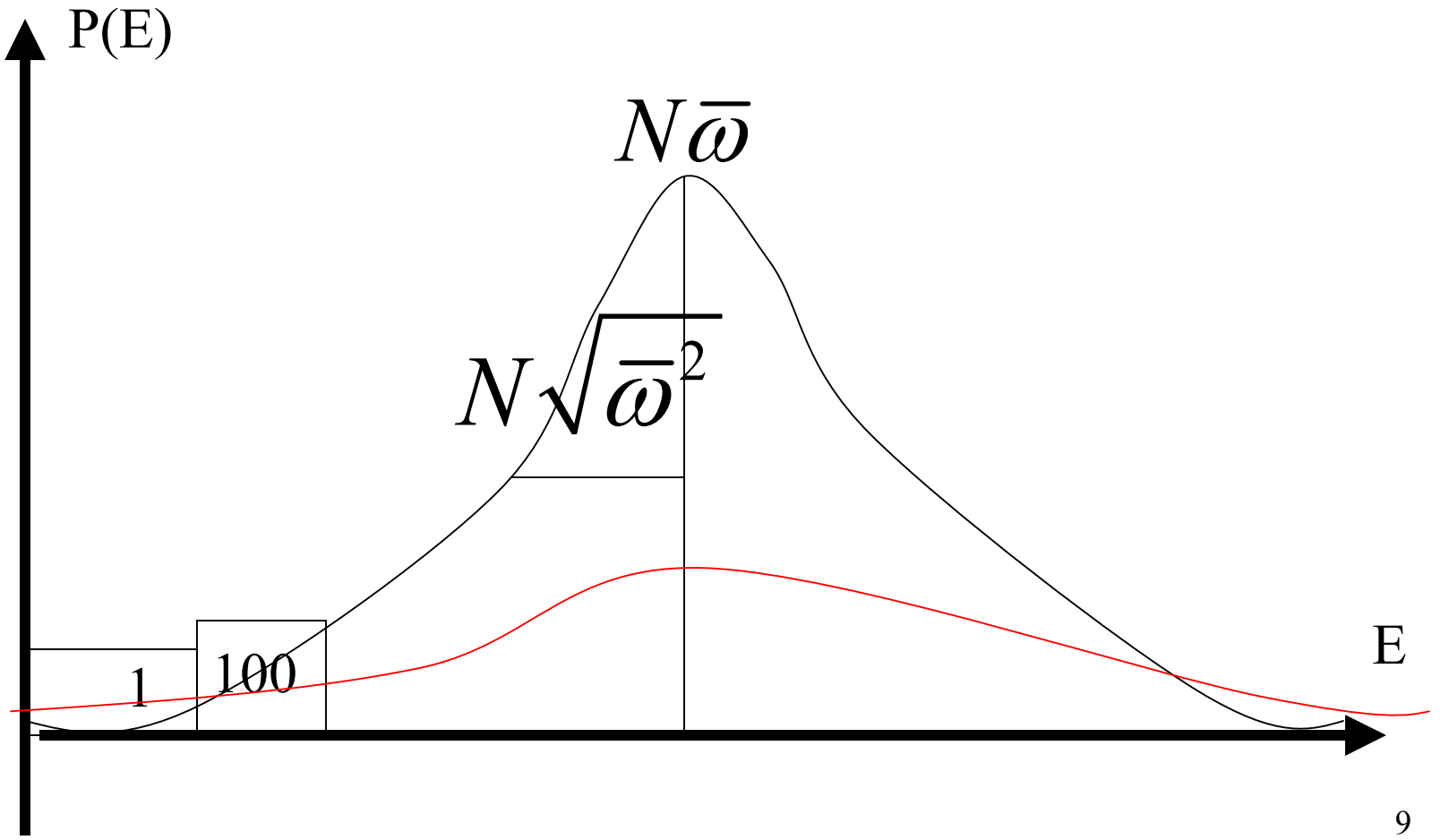
$$E = \sum_{i>j} h_{ij}\left(r_{ij} - r_c\right) u_{ij}\left(\alpha, \beta\right) \approx \sum_k \omega_k$$

$\left\langle \omega \right\rangle \equiv \bar{\omega}$ $\quad$ $\left\langle \omega^2 \right\rangle \equiv \bar{\omega}^2$ $\quad$ both $\bar{\omega}$ and $\bar{\omega}^2$ are finite

$$P(E)dE \simeq \left(\frac{2N\bar{\omega}^2}{\pi}\right) \exp\left[-\frac{E - N\bar{\omega}}{2N\bar{\omega}^2}\right] dE$$

Central limit theorem suggests a normal distribution to the energies assuming the individual interactions are independent random variables

The wider is the Gaussian the better separated (more stable) is the “native” structure



Good folders are when

$$\frac{E_{\min} - \langle E \rangle}{\sigma} \equiv Z$$

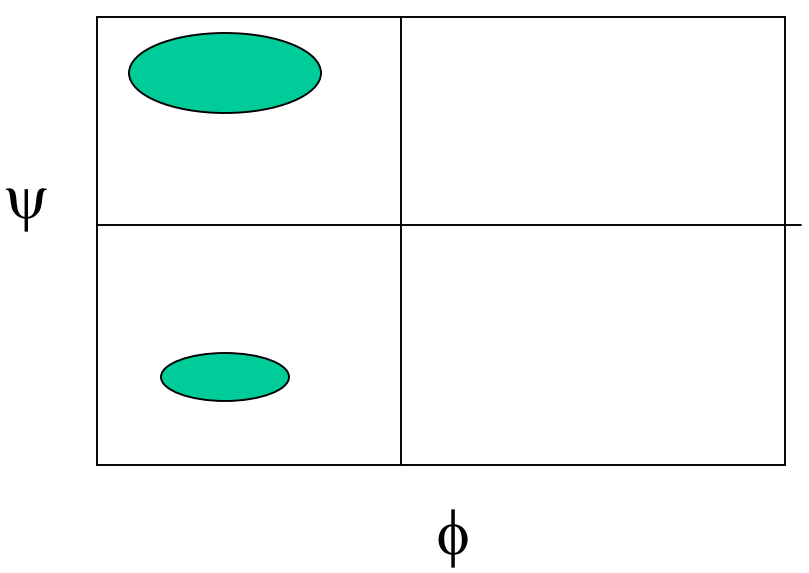
(definition of Z score – Z score determines if the annotation is significant)
Z scores is used in sequence alignments, in threading, in energy design and in ab-initio folding

The Random Energy Model (REM)

- The REM is general and not necessarily simple
  - Infinite number of "colors"
  - No specific model to be tested
  - Makes it possible to test a wide range of alternatives
  - Can we come up with a "toy" model that is (nevertheless) concrete?

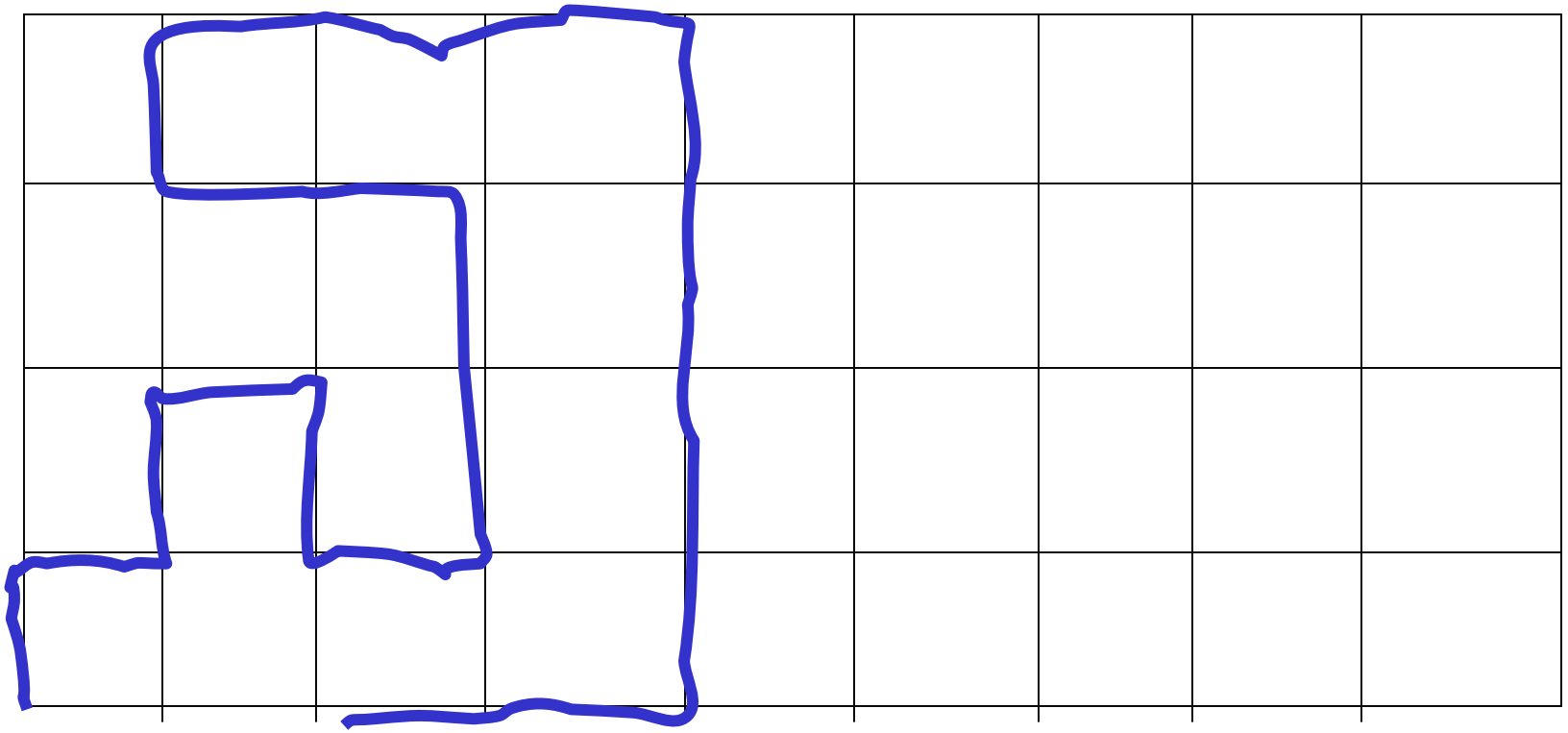
Lattice representation

- Protein chains are embedded in continuous space. Nevertheless, they approximately following discrete positions on a lattice.



$$\ldots - N \xrightarrow{\phi} C_\alpha \xrightarrow{\psi} C - N - \ldots$$

A lattice model (two dimensional cubic lattice, chain self avoiding)



Only two types of amino acids: H and P (Hydrophobic and Polar)

Hydrophobic in (hate water), Polar (hydrophilic) out

H=A,V,P,L,I,M,F,W,C,H,M

P=R,K,D,G,E,S,T,N,Q

$$u(\alpha,\beta) = \begin{array}{c} \\ H \\ P \end{array} \begin{array}{c} \begin{array}{cc} H & P \end{array} \\ \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix} \end{array}$$

Physically based. Minimum number of colors.

The stability of the structure is determined by the number of H-H contacts

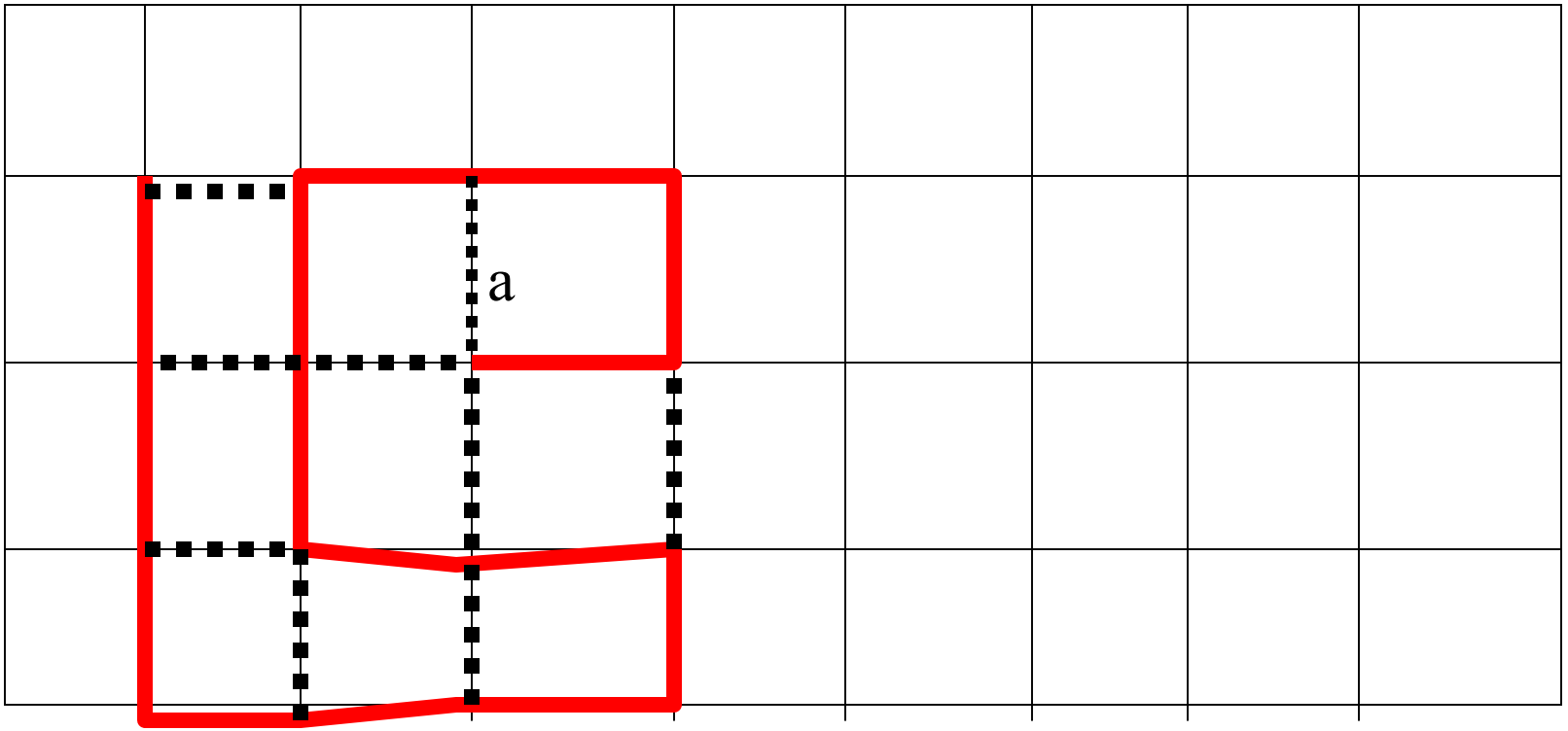
- No effect of amino acid sizes
- No effect of electrostatic interactions
- No secondary structure
- Much less than 20 types
- Nevertheless
  - No free parameters
  - Remarkable recognition capacity
  - Not clear if 20 types are required for stability

Two dimensional square lattice with an H/P polymer

- Contacts are defined between
  - two neighboring lattice points (distance “a”)
  - The two points are separated by more than two bonds
  - A lattice site will have between zero and three neighbors. Three neighbors are rare, why?
  - Consider a long polymer (length L) arranged in a compact configuration. What is the L dependence of monomers with one or two neighbors? What will be the situation in three dimensions?

A compact configuration in 2 dimensions

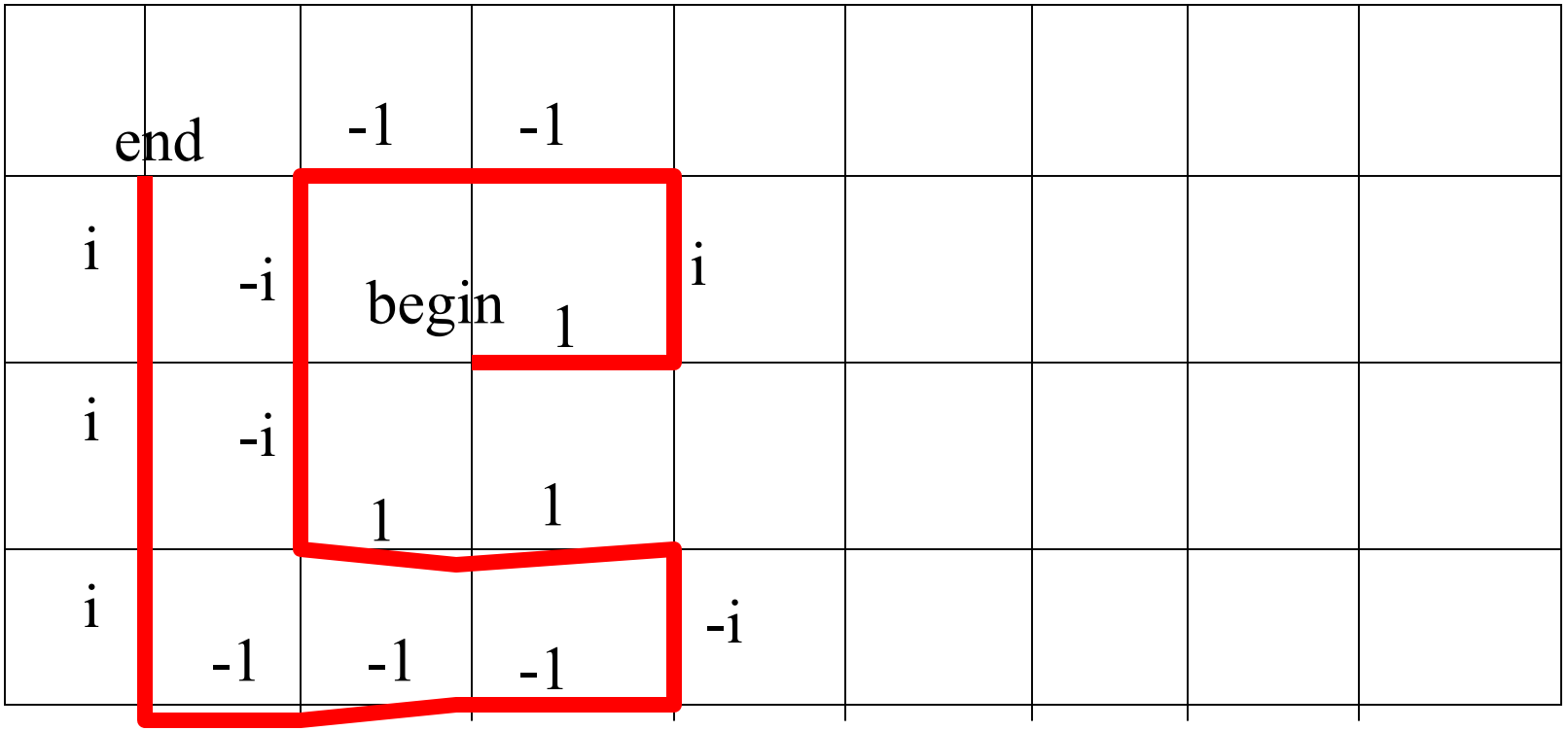
9 contacts, 16 amino acids



A representation of the two dimensional polymer

- Using pair of Cartesian coordinates, every amino acid (k) is represented by a pair of integers (n,m) indicating the lattice position. The polymer is presented by the triplets (k,n,m). This representation is useful in the calculation of contacts.
- Internal coordinate representation. Every bond between two sequential amino acid is presented by a complex number of four possible values - $1, -1, i, -i \quad i = \sqrt{-1}$. Useful for chain reconstruction and modeling the polymer motions. How can we determine if two chains are related by overall rotation and translation?

Example for internal coordinate representation



Metropolis algorithm for the "kinetics" of protein folding

- Determine initial configuration and initial temperature T
- Fix the direction of the first bond (why?)
- Start loop
- Assign the current protein energy to E(i)
- Choose a bond k at random
- Rotate bond k by +/- 90 degrees at random (multiply the bond value by +/-*i*)
- Construct a Cartesian representation of the chain on the lattice from the bond representation
- Reject the step if monomers overlap is detected
- Accept the step with a probability min(1,exp[-(E(i+1)-E(i))/T])
- If temperature is higher than zero decrease temperature and go to Start loop
- Otherwise stop

Is complete enumeration possible?

- The stochastic procedure (Metropolis algorithm) is very popular in three dimensions and more complex models of proteins chains. However in our case it is possible to directly consider all chain states for polymers than are not too long
- Suggests a rough upper and lower estimate for the number of chain conformations of chain lengths: 10, 50, 100. Are complete enumerations possible??

An example of a lattice in three dimension (diamond lattice)

Bound the number of alternative conformations On a diamond lattice as a function of polymer length