3.36pt

Protein Folding

Protein folding is the problem of deciding a protein's native state, i.e. a minimum energy, three dimensional fold, from its primary state, i.e. it's linear sequence of amino acids. This is one of the most well known and well studied problems in computational biology.

The underlying theoretical premises is due to Christian B. Anfinsen, who published experimental results in 1961 illustrating the Thermodynamic Hypothesis, which claims that a protein's native state minimizes free energy. There are numerous energy models that combine into a myriad of different energetic descriptions, leading to a host of approximation schemes to infer native states. There are even international competitions, see the Critical Assessment of protein Structure Prediction (CASP).

Energy Functions

Typical energy contributions are:

action	Energy	
stretching	$\text{Energy}^{\text{stretch}}$	$= \sum_{i} K_i^L (L_i - L_i^0)^2$
bending	$\text{Energy}^{\text{bend}}$	$= \sum_{i} K_i^\theta (\theta_i - \theta_i^0)^2$
twisting	$\text{Energy}^{\text{twist}}$	$= \sum_{i} K_i^\phi (1 - \cos(\omega_i))$
Electrostatic	$\text{Energy}^{\text{elec}}$	$= \sum_{i<j} K_{ij}^{\text{elec}} \frac{q_i q_j}{d_{ij}}$
Van der Waals	$\text{Energy}^{\text{vdw}}$	$= \sum_{i<j} K_{ij}^{\text{vdw}} \left(\left(\frac{d_{ij}^*}{d_{ij}} \right)^{12} - \alpha_{ij} \left(\frac{d_{ij}^*}{d_{ij}} \right)^{6} \right)$

Lattice Models

Lattice models reduce the complication of assembling an energy model by providing a structure for the folding process.

Structure The structure is a subset of $\mathbb{Z}^3$, i.e. a lattice of the integers. Each residue is assigned a unique location (i, j, k) with consecutive residues of the native state being one of $(i, j, k) \pm (1, 0, 0)$, $(i, j, k) \pm (0, 1, 0)$, or $(i, j, k) \pm (0, 0, 1)$.

Assumption Each residue is either hydrophobic (H) or hydrophilic (P). We encode H as 1 and P as 0.

Energy Energy is reduced to a single value $\varepsilon < 0$ for each HH pair between non-consecutive residues of the native state. These are called topological neighbors in the paper.

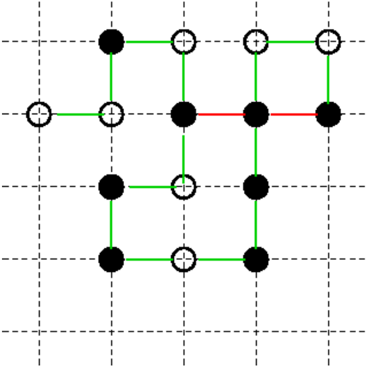
Goal Minimizing energy corresponds with maximizing the number of topological HH pairs.

A Simple Example

Suppose the native sequence is

$$\begin{array}{cccccccccccccccc} & P & P & H & P & H & P & H & H & P & H & H & H & P & P & H \\ s = & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{array}$$

One 2D lattice fold with two HH topological neighbors is



2D Models

We restrict our initial study to 2D models and seek a deterministic heuristic with guaranteed performance.

A sequence is decomposed into zero separators and blocks.

Zero Separators Denoted by z_i. Every sequence is assumed to start an end with a zero separator, and these are the only separators whose lengths are arbitrary (including zero). All other separators are possibly empty, consecutive strings of hydrophilic residues of even length (including zero).

Blocks Denoted by b_i. These are sequences of like 101000101 but not 11101. The latter is instead contains three blocks $b_1 = 1$, $b_2 = 1$ and $b_3 = 101$.

Sequence Decomposition

An illustrative block decomposition is

$$\begin{array}{ccccccccccccccccccc}
0 & \underline{1010001} & & \underline{1} & & \underline{1} & & \underline{1} & 00 & \underline{100000001} & & \underline{10101} & & \underline{1000101} & & \underline{101} & 00 & \underline{1} & \\
z_0 & b_1 & z_1 & b_2 & z_2 & b_3 & z_3 & b_4 & z_4 & b_5 & z_5 & b_6 & z_6 & b_7 & z_7 & b_8 & z_8 & b_9 & z_9 \\
 & y_1 & & x_1 & & y_2 & & x_2 & & y_3 & & x_3 & & y_4 & & x_4 & & y_5 & \\
\cline{1-10} \cline{11-19}
\multicolumn{10}{c}{B'} & \multicolumn{9}{c}{B''}
\end{array}$$

Notice that z_1, z_2, z_3, z_5, z_6, z_7, and z_9 are all empty zero block separators.

Blocks are alternatively labeled as y and x blocks.

B' and B'' are "super blocks," and we want to establish a division of super blocks that help identify minimum energy states.

Some Definitions

Definition

Positions on lattice $\mathcal{L} = \{1, 2, \ldots, m\}^2$ are neighbors, denoted by $(x, y) \sim (x', y')$, if $x = x' \pm 1$ with $y = y'$ or $x = x'$ with $y = y' \pm 1$.

Definition

A fold of sequence $s = s_1 s_2 \ldots s_n$ on lattice $\mathcal{L} = \{1, 2, \ldots, m\}^2$ is a one-to-one function $f : \{s_1, s_2, \ldots, s_n\} \to \mathcal{L} : s_i \mapsto (x_i, y_i)$ satisfying $f(s_i) \sim f(s_j)$ if $i = j \pm 1$.

Definition

Topological neighbors for fold f of sequence s are pairs (s_i, s_j) with $|i - j| \geq 2$ such that $f(s_i) \sim f(s_j)$.

Some Theory

Proposition

An HH topological pair must be separated by an even number or residues, i.e. if $s_i = s_j = 1$ with $|i - j| \geq 2$ and $|i - j|$ even, then $f(s_i) \not\sim f(s_j)$ for any fold f.

Proposition

There are no topological HH pairs within a block.

Proposition

Blocks b_i and b_j can provide topological HH pairs if and only if $|i - j|$ is odd.

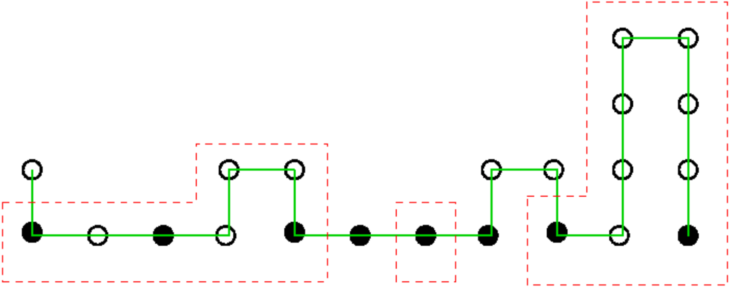
Proposition

Topological HH pairs can only arise between x and y blocks.

Folding a Super Block

We can fold a super block to maintain 'exposure' for either the x or y blocks. Suppose we want to fold B' from a few slides back to 'expose' the y blocks,

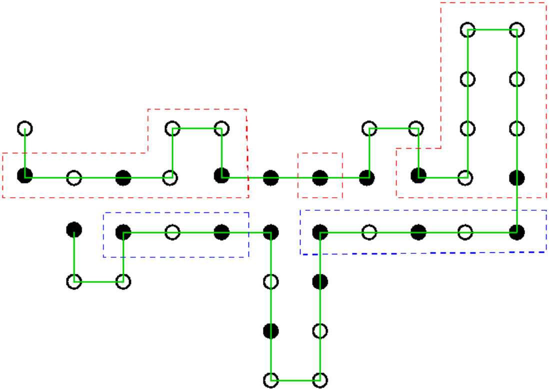
$$0 \quad \underbrace{1010001}_{y_1} \qquad \underbrace{1}_{x_1} \qquad \underbrace{1}_{y_2} \qquad \underbrace{1}_{x_2} \quad 00 \quad \underbrace{100000001}_{y_3}$$



Folding a Super Block

Here is an example of folding super blocks B' and B'', the first exposing y blocks (red), and the second exposing x blocks (blue).

$$B'' \quad \underset{x_3}{\underline{10101}} \quad \underset{y_4}{\underline{1000101}} \quad \underset{x_4}{\underline{101}} \quad 00 \quad \underset{y_5}{\underline{1}}$$



General Goals

Initial Goal Decide how to divide a sequence into two super blocks so that their joint 2D fold attempts to minimize energy. This is is Subroutine 1 of the paper. The idea is equate as well as possible the number of H residues in the x blocks in one of the super blocks with the number of H residues in the y blocks of the other super block.

Visualization Write code to generate a 2D fold from the two super blocks. You will want to consider folds different from those in the article. This code should also calculate the number of topological HH neighbors.

3D Extensions See if you can extend your algorithm to 3D folds.

Mathematics Prove the conjectures stated earlier.

Adaptations You might want to try Algorithm B if time remains (doubtful).

Data

Test your code on the 10 proteins listed in, *A test of lattice protein folding algorithms*, by K. Yue, K. Fiebig, P. Thomas, H. Chan, E. Shaknovich, and K. Dill, PNAS, vol. 92, 1995, pp 325-329.