Protein Folding: Lattice Models

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

December 2, 2010

Outline

1. Outline

2. Protein Folding
    - The HP-Model

3. Folding Algorithms for the 2D Square Lattice
    - Approximation Algorithms for Side-Chains Models

Computational Protein Folding Model

“The protein folding problem is three different problems: the folding code – the thermodynamic question of how a native structure results from the interatomic forces acting on an amino acid sequence; protein structure prediction – the computational problem of how to predict the native structure of a protein from its amino acid sequence; and the folding speed (Levinthal’s paradox) – the kinetic question of how a protein can fold so fast ... Ken Dill 2007

Computational Protein Folding Model

... Current knowledge of the folding codes is sufficient to guide the successful design of new proteins and new materials. Current computer algorithms are now predicting the native structures of small simple proteins remarkable accurately, contributing to drug discovery and proteomics. Even once intractable Levinthal puzzle now seems to have a very simple answer ...” Ken Dill 2007

Computational Protein Folding Model

The protein folding problem is in fact a collection of fundamental problems focused on the questions, "What is the folding code?" and "What is the folding mechanism?" and " . . . the second, more visible to the public, side of the 'holy grail' of protein folding – prediction of protein conformation. The "folding code" concerns how the "tertiary structure and folding pathway of a protein are encoded in its amino acid sequence...[it] is not predominantly localized in short windows of the amino acid sequence . . . [it] resigns mainly in global patterns of interactions, which are nonlocal, and arise from the arrangements of polar and non-polar monomers in the sequence".

Computational Protein Folding Model

"The failure of protein-folding processes, both within cells (in vivo) and within test tubes or industrial vats (in vitro), causes serious difficulties both for biomedical research and for biotechnology industry. Protein chains that fail to fold properly aggregate into an insoluble and inactive state... There is increased recognition that some human diseases are associated with aberrations or defects in protein chain folding. These include Alzheimer's and Huntington's and cystic fibrosis." Jonathan King 2002

Computational Protein Folding Model

“Understanding the mechanism of protein folding is often called the “second half” of genetics. Computational approaches have been instrumental in the efforts. Simplified models have been applied to understand the physical principles governing the folding processes and will continue to play important roles in the endeavor.”
Peter Kollman 2001

Computational Protein Folding Model

“[W]e take as our premise that proteins are chain molecules that have specific monomer sequences and are driven to fold mainly by nonlocal interactions subject to steric constraints. There is currently no accurate analytical theory that can account for chain connectivity, excluded volume in the compact states, and specific sequences of monomer units. Simple exact models have been developed to explore such properties.”

Computational Protein Folding Model

“Folding is an intrinsically statistical phenomenon and no conclusion can be derived from a single folding or unfolding trajectory. . . . Lattice and other simplified analytical models are the statistical mechanician’s contribution to the protein folding . . . their intimate connection with statistical mechanics . . . is very important as it often allows us to compare simulation with statistical-mechanical analytical theories.” Eugene Shakhnovich 1996

The HP-Model: Hydrophobic and Hydrophilic amino acids

- In 1985, Ken Dill proposed the hydrophobic-hydrophilic (HP) model, which has been subjected to a huge amount of literature due to its fundamental role in protein folding modeling.
- The model captures the fact that native protein folds tend to form very compact cores driven by dominant hydrophobic interactions.
- Each amino acid is classified either as hydrophobic (H) or hydrophilic (P) and two hydrophobic amino acids are said to be in *contact* if they are adjacent in the fold but nonadjacent in the primary sequence.

The HP-Model: H-H contacts

- Since the goal is the formation of highly compact hydrophobic cores, the optimization function is to maximize the number of contacts between hydrophobic atoms (H-H contacts).
- To phrase the problem as an energy-minimization problem, the energy function is the negative of the number of hydrophobic contacts of the fold.
- Two examples of protein folds in the linear-chain and side-chain lattice HP-models are presented in Figures 2 and 5.

The HP-Model: side-chain models

- Models represent the protein sequence as a linear chain, perhaps with explicit side-chains branching from the linear backbone.

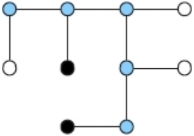


Figure: A fold of a protein in the 2D square side-chain HP-model making one contact

The HP-Model: Lattice and off-lattice models

- In *lattice models*, a fold of a protein sequence is defined by placing the amino acids on lattice nodes and the protein chain as a self-avoiding path on the lattice;
- in *off-lattice models*, the placement of the protein is in 3D space, with the only restriction being the self-avoidance of the backbone and of the branching side-chains.

The HP-Model: Lattice and off-lattice models

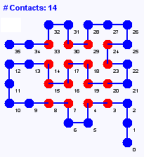


Figure: Optimal fold of a protein of length 36 (red is hydrophobic and blue is hydrophilic)

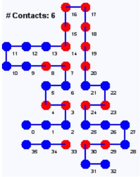


Figure: A fold constructed by the Hart-Istrail algorithm

Side-Chain Models

To increase the accuracy of protein prediction methods, it is desirable that extended models take into account the structure of the protein as a backbone formed by a set of successive peptide bonds, together with attached side chains.

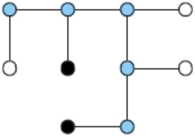


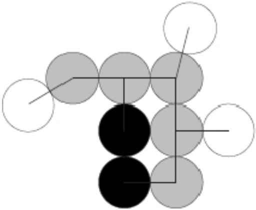
Figure: A fold of a protein in the 2D square side-chain HP-model making one contact

Side-Chain Models

- The side-chain lattice models analyzed represent the folding of proteins as "branched combs."
- In the side-chain model, the backbone of the protein is represented by a linear sequence of backbone nodes (as in the HP-model, except that these nodes are not labeled with amino acids), and connected to each backbone node is a side chain (an edge with one end a backbone node and the other end representing the amino acid) representing an amino acid (labeled either hydrophobic or hydrophilic) (Figures 5) .
- A conformation of the protein is an embedding in a "self-avoiding manner" of the backbone path into the lattice with side-chain edges mapped to adjacent lattice edges such that no lattice point is occupied by more than one backbone node or side-chain node. As before, the energy of a conformation is the number of hydrophobic-hydrophilic contacts between amino acids.

Off-Lattice Models

Off-Lattice Models. An important question to address in studying lattice models is whether or not the algorithms for these models can be generalized to algorithms for off-lattice models. In 1997, Hart and Istrail introduced an off-lattice model called the Tangent Spheres side-chain HP-model (HP-TSSC). In this model, adjacent backbone and side-chain molecules are represented by identical spheres in 3D space that are tangent. Side chains are labeled hydrophobic or hydrophilic and the energy of a conformation is the number of hydrophobic-hydrophobic tangent spheres.



Hart-Istrail 2D Algorithm

Theorem (Hart-Istrail 1995)

There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 2D square HP-model within $\frac{1}{4}$ *(0.25) of optimal.*

Hart-Istrail 2D Square Lattice Algorithm

For every protein sequence S in the linear chain HP-model, consider the following:

1. $OPT_{2D}(S)$ = the maximal number of contacts of any fold of S on the 2D square lattice; $OPT_{3D}(S)$ = the maximal number of contacts of any fold of S on the 3D cubic lattice;
2.
$$\mathcal{C}_{2D}(S) = 2\min\{\mathcal{O}(S), \mathcal{E}(S)\}$$
$$\mathcal{C}_{3D}(S) = 4(\min\{\mathcal{O}(S), \mathcal{E}(S)\}) + 2$$

Theorem (Hart-Istrail 1995)

For every protein sequence S in the linear chain HP-model:

1. $OPT_{2D}(s) \leq \mathcal{C}_{2D}(S)$
2. $OPT_{3D}(s) \leq \mathcal{C}_{3D}(S)$

Hart-Istrail 2D Square Lattice Algorithm

Therefore, every 2D algorithm that constructs folds achieving a fraction of α of the $\mathcal{C}_{2D}(S)$ contacts is an approximation algorithm with ratio α; it is then guaranteed to achieve at least α of the optimal number of contacts. Similarly for the 3D case.

Hart-Istrail 2D Square Lattice Algorithm

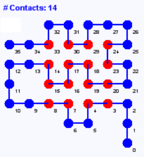


Figure: Optimal fold of a protein of length 36 (gray is hydrophobic and black is hydrophilic)

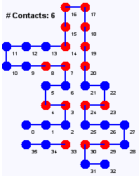


Figure: A fold constructed by the Hart-Istrail algorithm

Newman 2D Lattice Algorithm

Theorem (Newman 2002)

There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 2D square HP-model within $\frac{1}{3}$ (0.33) of optimal.

Newman 2D Lattice Algorithm

Theorem (Newman 2002)

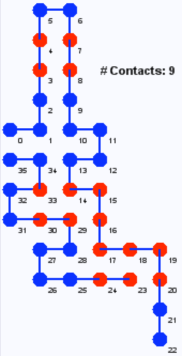


Figure: A fold constructed by the Newman algorithm

Hart-Istrail 3D Cubic Lattice Algorithm

Theorem (Hart-Istrail 1995)

There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the 3D cubic HP-model within $\frac{3}{8}$ *(0.38) of optimal.*

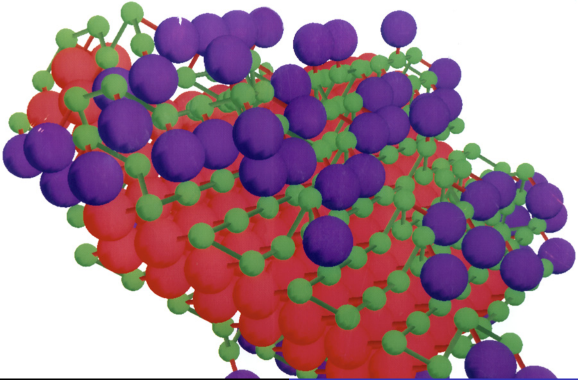
Newman 2D Lattice Algorithm

NAME	YEAR	LATTICE TYPE	MODEL TYPE	APPROX. RATIO
Jiang-Zhu	2005	2D Hexagonal Lattice	Side-Chains	$\frac{1}{6}$ (0.17)
Hart-Istrail	1995	2D Square Lattice	Linear Chains	$\frac{1}{4}$ (0.25)
Mauri-Pavesi *et al*	1999	2D Square Lattice	Linear Chains	$\frac{1}{4}$ (0.25)
Newman	2002	2D Square Lattice	Linear Chains	$\frac{1}{3}$ (0.33)
Newman-Ruhl	2004	3D Cubic Lattice	Linear Chains	$\frac{3}{8}$ (0.3750)
Hart-Istrail	1995	3D Cubic Lattice	Linear Chains	$\frac{3}{8}$ (0.38)
Batzoglou-Decatur	1996	2D Triangular Lattice	Linear Chains	$\frac{1}{2}$ (0.5)
Agarwala-Batzoglou *et al*	1997	2D Triangular Lattice	Linear Chains	$\frac{6}{11}$ (0.55)
Agarwala-Batzoglou *et al*	1997	3D Triangular Lattice	Linear Chains	$\frac{44}{75}$ (0.59)
Batzoglou-Decatur	1996	3D Triangular Lattice	Linear Chains	$\frac{3}{5}$ (0.60)
Bokenhauer-Bongartz	2007	3D Cubic w/ Diagonals	Linear Chains	$\frac{5}{8}$ (0.62)
Heun	2003	2D Square w/ Diagonals	Linear Chains	$\frac{59}{70}$ (0.84)
Hart-Istrail	1997	FCC Lattice	Side-Chains	$\frac{31}{36}$ (0.86)
Hart-Istrail	1997		Off-Lattice	$\frac{31}{36}$ (0.86)

Hart-Istrail Side-Chain Face-Centered Cubic Lattice Algorithm

Theorem (Hart-Istrail 1997)

There is a linear-time approximation algorithm that folds an arbitrary HP-protein sequence in the extended 3D FCC lattice HP-model within $\frac{31}{36}$ (0.86) of optimal.



Hart-Istrail Side-Chain Face-Centered Cubic Lattice Algorithm

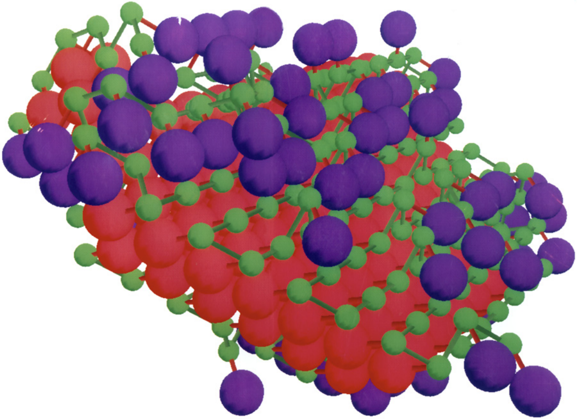


Figure: A PDB protein represented in the HP-side chain model, folded near optimally (98%) on the FCC lattice. Red is hydrophobic, Pink is hydrophilic, Green is backbone.