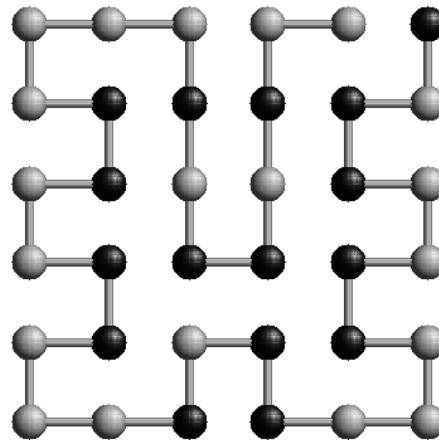


Protein Design in the 2D HP Model

A Monte-Carlo Iterative Design Approach



Reza Lotun and Camilo Rostoker
{rlotun,rostokec}@cs.ubc.ca
Department of Computer Science, UBC

Presentation Outline

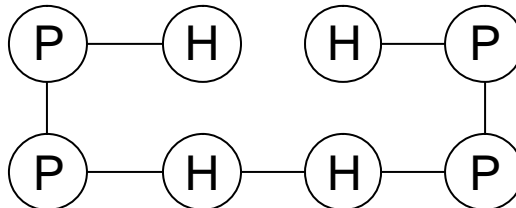
1. Review of proteins and the central dogma
2. Introduction to the 2D HP Model
3. Overview of Existing Approaches
4. The Iterative Design Approach
5. Our Enhancements to the Iterative Design
6. Experiments & Analysis
7. Conclusion and Future Work

Brief Protein Review

- Recall the Central Dogma: DNA→RNA→Proteins
- So far we've sequenced the human genome. Now we have to understand proteins.
- One thing we do know about proteins: **STRUCTURE** is all important
- **Anfinsen's Experiment:** in 1950, he showed the link between the amino acid sequence and protein 3D structure → *Sequence specifies Conformation*
- **Levinthal's Paradox:** Nature has somehow optimized the protein folding process

The 2D HP Model

- First proposed by Dill in 1989
- Motivation: several popular views of hydrophobic and polar amino acids
 - The Hydrophobic Force is the #1 driving force for protein folding
 - Native conformations are compact with dense hydrophobic cores with a shell of polar amino acids exposed to the solvent
- The 20 amino acids are reduced to hydrophobic (H) and polar (P)
- The protein structure is reduced to a 2D self-avoiding walk on a lattice (Conformation)



Justification of 2D HP Model

- Why the 2D HP Model anyways?
 - Simplified view of both sequences and conformations → reduced complexity
 - Exact enumeration is possible
 - The Lattice/SAW models are well-studied problems
- What can we learn about proteins from this 2D HP model?
 - General patterns of hydrophobic interactions
 - Gain insight into complex protein structure parameters

Problem Definition

- **Problem Definition:** given a conformation C with N residues, and a set of possible solution sequences SET , find a sequence S of length N such that:
 1. $energy(S,C)$ is the lowest for all s in S
 2. There is no other conformation C^* of length N such that $energy(S,C^*) < energy(S,C)$

where $energy(S,C)$ is defined as the # of H-H contacts in the conformation

- **Native Conformation:** A conformation C that satisfies the above constraints is called the *Native Conformation of S*
- **Uniqueness:** Every sequence has *exactly one* Native Conformation; however, a conformation may be Native for more than one sequence

Formal Definitions

From a thermodynamic perspective, given some energy function which depends on the interactions between H and P amino acids, the conditional probability of a sequence s folding into conformation G_0 at temperature T is denoted by (Boltzmann Distribution):

$$P(\Gamma_0 | \sigma) = \frac{1}{Z(\sigma)} \exp[-E(\Gamma_0, \sigma)/T]$$

Where Z is the *partition function*

$$Z(\sigma) = \sum_{\Gamma} \exp[-E(\Gamma, \sigma)/T]$$

We also define the *Free Energy* as

$$F(\sigma) = -T \ln Z(\sigma)$$

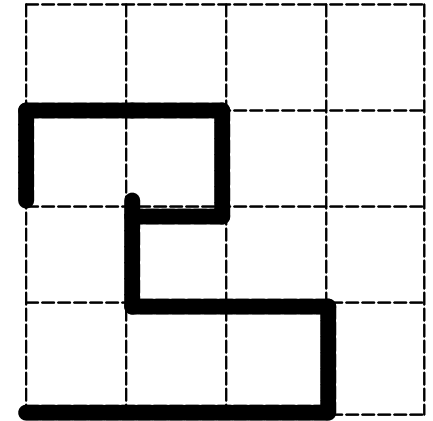
SAW Representation

1. **Coordinate Representation:** treat each site in the SAW as point in a Cartesian plane \rightarrow a whole walk is a list of coordinates

2. **Quad Notation:** treat each segment as a *global move* from the alphabet $\{U, D, R, L\}$ (Up/North, D/South, R/East, L/West).

3. **Ternary Representation:** treat each segment as a *local move* from the alphabet $\{F, R, L\}$ (Forward, Right, Left), which can be encoded by a ternary string $\{L=1, F=2, R=3\}$.

3.5 **Balanced Ternary:** Take the ternary representation and subtract 2 to get alphabet $\{L=-1, F=0, R=1\}$.
 \rightarrow can reverse a SAW by multiplying by -1
 \rightarrow * **SAW equivalency:** two SAW's that are the same will have the same or reversed balanced ternary representation.



(0,0), (1,0), (2,0),
 (3,0), (3,1), (2,1),...,
 (0,2)

RRRULLURULLD

FFFLLFRLLFL

222112331121

000-1-1011-1-10-1

Criteria for Good Sequence Design

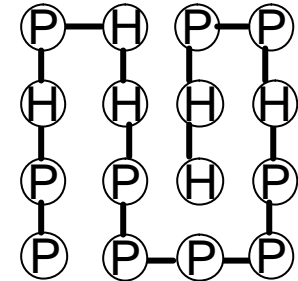
1. The sequence should have the desired conformation when it is in its ground state
→ **positive design**
2. There should be no ground state degeneracy, that is, the sequence should always fold into the same conformation
→ **negative design**
3. There should be a large gap in the energy of the ground state and the energies of higher states
→ **sequence stability**

Degenerate Sequences

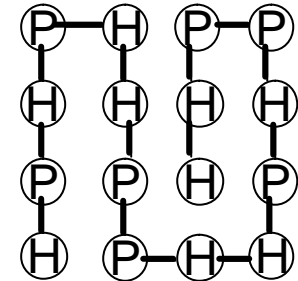
Example: consider the first conformation on the right

Sequence 1: PPHPHPPPPPHPPHH

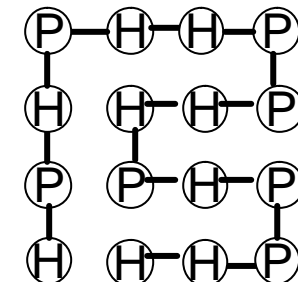
Sequence 2: HPHPHHPPHHPHPPHH



Sequence 1 is a designing sequence which has energy = -3



Sequence 2 has energy = -4 and appears to be a designing sequence...but we can find another conformation on which it has energy -6



The sequence is **degenerate**, since it can degenerate from one conformation to another

→ **unstable, not a solution sequence**

Existing Approaches

1. Free energy is ignored or set to a constant. The problem then simply becomes one of minimizing E.

$$P(\Gamma_0 | \sigma) = \frac{1}{Z(\sigma)} \exp[-E(\Gamma_0, \sigma)/T]$$

2. A high temperature expansion is made on F

→ define $F = E(G_0, s_i) - F(s_i)$
(a value we want to minimize)

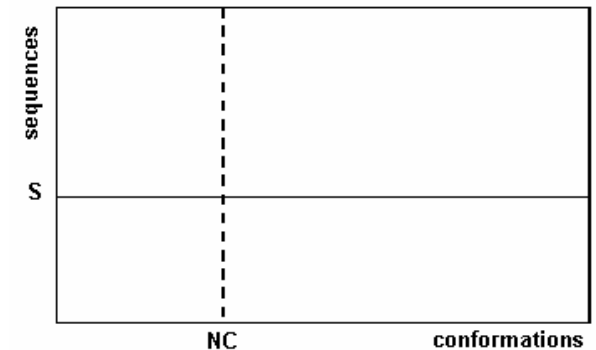
→ expand F with Taylor approximation
keep the lowest order term and
ignore constants

$$Z(\sigma) = \sum_{\Gamma} \exp[-E(\Gamma, \sigma)/T]$$

$$F(\sigma) = -T \ln Z(\sigma)$$

3. A **Monte Carlo** (MC) algorithm is used to estimate F (sample search space)

1. Fixing the sequence (nested MC)
2. Both sequence and conformation updated simultaneously.



Iterative Design Approach

- First proposed by Micheletti, Rossi & Maritan in 1999
- **Main Idea:** Iteratively select the best sequence S using an approximate Free Energy sum, then check that it does not have lower energy on some other conformation
- Addresses problem of degenerate sequences by maintaining a set of *decoy conformations*
- Small number of optimal decoys contribute majority of energy to free energy sum
- Decoys are therefore used as the basis to calculate an approximate free energy of a sequence S (instead of summing over ALL possible conformations)
 - Major increase in efficiency
 - Weeds out putative solutions before costly step of searching conformation space

Pseudo-code of Existing Algorithm

GIVEN: a native conformation NC

INITIALIZATION:

Generate a set of possible solution sequences
Initialize a set of decoy structures (can be empty)

REPEAT UNTIL TERMINATION CRITERIA IS MET

Exhaustively search sequence space to find a sequence s that minimize:

$$\beta[\text{energy}(s, \text{NC}) - \text{FE}(s, \beta)] < \ln 2$$

Find set of conformation(s) C on which s has lowest energy

IF for any c in C, $\text{energy}(s, c) < \text{energy}(s, \text{NC})$,

s is not a solution

Add c to the set of decoys

ELSE

add s to the solution set

TERMINATION CRITERIA

No more sequences that satisfy F.E. sum condition; or

A sufficient number of solutions has been found

Our Enhancements

Main Idea: Improve on exhaustive steps

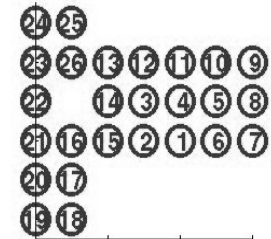
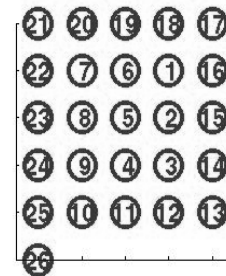
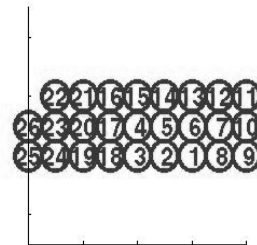
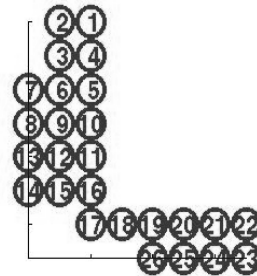
1. Non-exhaustive searching of sequence space using Simulated Annealing
2. Non-exhaustive searching of conformation space using Monte Carlo sampling
3. Pre-populating decoy set

Sampling Compact Conformations

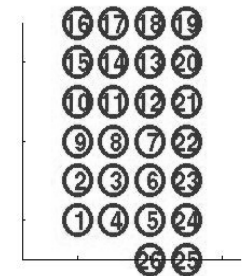
Interacting Growth Walk (IGW)
algorithm of Narasimhan et al.

$$p_m(r_j) = \frac{\exp(-\mathbf{b}n_{NN}^m(j)\mathbf{e}_0)}{\sum_{m=1}^{z_j} \exp(-\mathbf{b}n_{NN}^m(j)\mathbf{e}_0)}$$

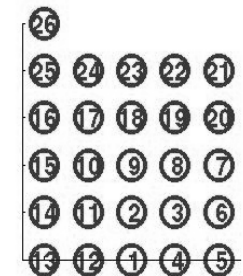
Give higher weight to those sites on the SAW that have fewer non-bonded nearest neighbors, resulting in a more compact SAW.



tbeta = 1



tbeta = 5



tbeta = 10

Fast Searching Through Sequence Space Using Simulated Annealing

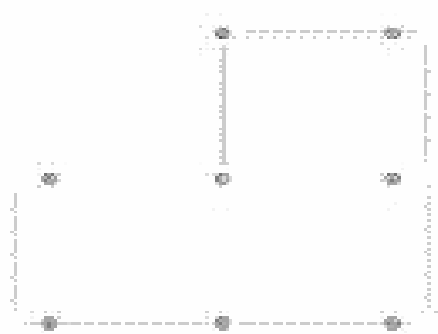
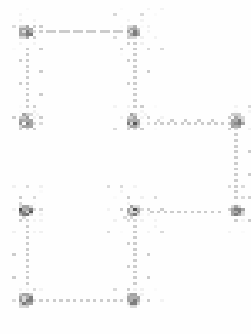
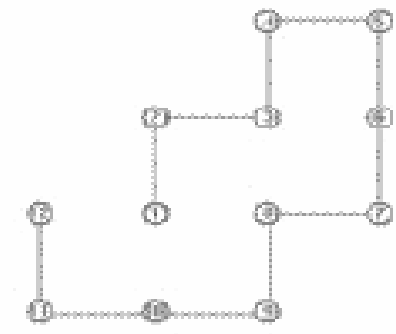
- A stochastic search strategy that is relatively efficient and is good for escaping from local minima
- Objective function is $\beta[\text{energy}(s, NC) - FE(s, \beta)] < \ln 2$
- Candidate solutions generated by perturbing sequence
 - Randomly switch between 1 and N individual positions
 - Replacing an entire segment with another randomly chosen sequence
- Metropolis step used for choosing a candidate solution
 - Accept new candidate solution if it has lower energy than original sequence, or accept with probability $p = -\Delta \text{Energy} / T$

Pre-populating Decoy Set

- Most significant contribution to the F.E. of a given sequence comes from “competing structures”
 - accuracy of the free energy approximation is dependant on quality of decoys
- **Idea:** Pre-populate the decoy set with “quality” decoys to increase accuracy of F.E. approximation earlier in iterations
- **Problem:** How do we create “quality” decoys?
- **Solution:** Use modified version of the Pivot Method
- **Pivot Method:** Choose at random a pivot, k , along the walk. Choose some symmetry transformation at random, and apply transformation to $k+1..N$. Check if self intersection occurs.

Experiments & Analysis

- Main Testing Goal:
 - Analyze tradeoff between sample size and accuracy
- Testing strategy and parameter selection
 - Sequence/conformation sizes
 - Conformation sample size
 - Conformation “compactness”
 - Simulated Annealing parameters

Test Case #1:	Test Case #2:	Test Case #3:
N = 8 URDDL U 2 solutions	N = 10 URDRDL DLU 3 solutions	N = 12 URURDDL DLU 4 solutions
		
Solutions/energy: HPPHPHPH 3 HHPHPHPH 3	Solutions/energy: HPPHPDPHPH 4 HHPHPDPHPH 4 HPPHPDPHPH 4	Solutions/energy: HPPHPDPHPHPH 5 HHHPDPHPDPHPH 5 HPHHPDPHPDPHPH 5 HHHPDPHPDPHPH 5

N	Sampling Exponent	# of samples taken	# iterations to finish	MI	MA	Actual # of solutions	# solutions returned	# of correct solutions
10	1.7	501	11	20	100	3	10	2
10	1.7	501	11	50	30	3	10	3
10	2.3	1995	13	20	100	3	10	3
10	2.3	1995	12	50	30	3	10	3

Conclusion

- Protein Design is complex, many complicating factors. Idea → Make assumptions, simplify it! → Still hard!
- Principal contribution: the ability to specify the degree of accuracy desired in the solution set (efficiency tradeoff)
- Remove exhaustive steps, replace with approximate searching using Monte Carlo and simulated annealing
- Pre-populating decoy set
- Guiding principle: designing real-life proteins that are much longer and more complicated requires non-exhaustive techniques

Future Work

- More testing on larger conformation lengths
- Optimize parameters
- More appropriate sequence search method
 - Stochastic Local Search techniques like GA, PSO, ACO, etc.
- Implementation details
 - Memory problems in MATLAB
 - Code Optimization
- From 2D → 3D
- Larger amino acid classes

References

Images

1. Protein structure images:
<http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookCHEM2.htm>
2. Sequence/Conformation energy plot:
A. Irback, C. Peterson, F. Pottharst, and E. Sandelin: Monte Carlo procedure for protein design. Physical Review E, Vol 58. Num 5, pp. 5249-5252, 1998.
3. 2D energy landscape cross-section:
<http://www.nada.kth.se/~asa/hopfield.html>
4. 3D energy landscape:
<http://www.press.uillinois.edu/epub/books/brown/ch7.html>

Papers

1. R. Lotun, C. Rostoker: "CPSC 545 Project Proposal: An Improved Iterative Protein Design Strategy", 2004
2. R. Lotun, C. Rostoker: "CPSC 545 Project Progress Report: An Improved Iterative Protein Design Strategy", 2004
3. Reinhard Scheimann, Exact Enumeration of 3D Lattice Proteins. (University of Leipzig, Diploma Thesis).
4. Erik Sandelin, Thermodynamics of Protein Folding and Design (Lund University, Thesis).
<http://www.rpc.msoe.edu/cbm2/images/gfp/gfp1-1.jpg> (Center for Biomolecular Modelling).

References Cont'd

3. Reinhard Scheimann, Exact Enumeration of 3D Lattice Proteins. (University of Leipzig, Diploma Thesis).
4. Erik Sandelin, Thermodynamics of Protein Folding and Design (Lund University, Thesis).
<http://www.rpc.msoe.edu/cbm2/images/gfp/gfp1-1.jpg> (Center for Biomolecular Modelling).
5. Lingso and Pederson

EOF

Implementation Environment

- Coded in MATLAB
 - Easy prototyping due to:
 - Fast matrix calculations
 - high-level programming
 - data-visualization tools
 - portability
 - Memory limitations
- Testing system:
 - Sun Fire V880 with 8 x 1.2 GHz UltraSPARC III processors
 - 32GB RAM
 - SunOS 5.9

Solution Refinement Step

- At the start, there are a small number of decoys (as few as one) → less accurate approximation
- Degenerate sequence may be selected as solution
- As iterations continue, more conformations are added to decoy set → better approximation
- Idea: Re-evaluate the current set of solution sequences for false-positives
- Provides extra layer of degeneracy check, and has negligible computational overhead