Protein Landscape Analysis in the Grand Canonical Model

James Aspnes, Julia Hartling, Ming-Yang Kao, Junhyong Kim, Gauri Shah
Outline

• Introduction

• Grand Canonical Model

• Tools
  – Linear Programming
  – Network Flow
  – A Compact Representation of all Minimum Cuts

• Algorithmic Results

• Computational Hardness Results

• Experimental Results
Introduction

Basic Question: Find a *fittest sequence* given a target protein structure

- Heuristic to search space of sequences  
  Sun et al. 1995

- Computational Tractability: open  
  Hart 1997

- Polynomial time solution  
  Kleinberg 1999
Grand Canonical Model (Sun et al)

Given: Target Structure with n residues
Design: Sequence x

\( x_i = \text{Hydrophobic(H)} \) or Polar (P)

Fitness Function:

\[
\phi(S) = \alpha \sum_{i < j-2} g(d_{ij}) + \beta \sum_{i \in S_H} s_i
\]

- \( s_i \): Solvent-accessible surface of residue \( i \) (Å)
- \( d_{ij} \): Distance between residues \( i \) and \( j \) (Å)

- Low solvent-accessible area (\( s_i \))
- Compact hydrophobic core
- \( \alpha < 0 \) and \( \beta > 0 \)

\[
g = \begin{cases} 
\frac{1}{1+\exp(d_{ij}-6.5)} & \text{when } d_{ij} \leq 6.5 \\
0 & \text{when } d_{ij} > 6.5 
\end{cases}
\]
1. Linear Programming

\[ \text{i-th residue : 0-1 variable } x_i \]

\[ \Phi(S) = \alpha \sum_{i<j-2 \atop i,j \in S_H} g(d_{ij}) + \beta \sum_{i \in S_H} s_i \]

\[ \Phi(x) = - \sum_{i<j-2 \atop i,j} a_{ij} x_i x_j + \sum b_i x_i \]

Let \( \Delta = \# \) of terms in the fitness function

minimize

\[ g(x, y) = - \sum a_{ij} y_{ij} + \sum b_i x_i \]

subject to

\[ 0 \leq x_i \leq 1 \quad \forall i \]

\[ 0 \leq y_{ij} \leq 1 \]

\[ y_{ij} \leq x_i \]

\[ y_{ij} \leq x_j \]

\[ \forall i, j : a_{ij} \neq 0 \]

- Linear (not quadratic)
- \( y_{ij} = \min(x_i, x_j) = x_i x_j \)
- Unimodular matrix \( \rightarrow \) Integral solutions
Theorem 1: \( \hat{x} : 0\text{-}1 \) vector. Can find fittest \( x \) with Minimum Weighted Hamming Distance 
\[
\sum_i w_i |x_i - \hat{x}_i|
\]

Proof:

\[
0 < \epsilon < \frac{1}{Wd(n+1)}; \ W \geq \max |w_i|
\]

Minimize \( f_\epsilon(x) = f(x) + \sum_i \epsilon w_i |x_i - \hat{x}_i| \)

Suppose \( x \) minimizes \( f(x) \).

\[
\begin{align*}
    f_\epsilon(x) & \leq f(x) + \frac{Wn}{Wd(n+1)} \\
    & < f(x) + \frac{1}{d} \\
    & \leq f(x') \\
    & \leq f_\epsilon(x')
\end{align*}
\]

\( x \) minimizes \( f_\epsilon(x) \) \( \Rightarrow \) \( x \) minimizes \( f(x) \)

Running Time: \( O(\Delta^2 \log \Delta) \)
2. Network Flow (Kleinberg 99)
Capacity of Min-Cut $= \Phi(S) + B$

$$\Phi(S) = \alpha \sum_{i<j-2} g(d_{ij}) + \beta \sum_{i \in S_H} s_i$$

Goldberg-Tarjan Min Cut: $O(VE \log(V^2/E))$
Compact Min Cut Representation

Picard-Queyranne 1980
1-1 correspondence: Min Cuts and Ideals
Space of all Fittest Sequences

**Theorem 2** Can enumerate all 0-1 \( f \)-minimizing vectors \( x \) in \( O(n) \) time per vector.

**Proof:** Steiner 1986 : Enumerates the ideals in \( O(n) \) time per ideal \( \square \)

**Theorem 3** Can find diameter \( (k) \) in Hamming distance of the set of 0-1 vectors \( x \) minimizing \( f(x) \) in \( O(n) \) time.

\[
k = |\rho^{-1}(V(\widehat{G}^f_{s,t}))| = 4
\]
Fittest Sequence for k functions

**Theorem 4** Given k fitness functions, determine if no simultaneous solution exists, or construct a graph which represents all possible solutions.
Adjacency of fittest sequences

\[ \Lambda = \text{Set of allowable mutations} \]

Downward closed system of subsets of \( \{1, \ldots, n\} \)

e.g. \( \{\{1, 3, 5\}, \{1\}, \{3\}, \{5\}, \{1, 3\}, \{3, 5\}, \{1, 5\}\} \)

HPPPHHPPHP
HPHPPHPHPHP
PPHPPHPHPHP

RESIDUES \( \xrightarrow{\text{ }} \) PQ GRAPH

Smallest \( \Lambda = \rho^{-1}(I \triangle I') \)

\( I : \text{Ideal of } x \)
\( I' : \text{Ideal of } x' \)
Necessary:

Sufficient:
Enumerating sub-optimal sequences

Lawler’s method (1972)
Best-First Search

At each step:
- Pick f-minimizing vector
- Remove and replace by \( n \) pairs
- Update priority queue

Total cost: \( O(n\Delta^2 \log \Delta) \) per value
Structure of space of all fittest seq

g


f
Computational Hardness Results

**Theorem 5** $\Omega_f = \text{set of all } x \text{ that minimize } f(x)$. \#P-complete problems:

1. **Computing the size of $\Omega_f$.**

2. **Counting the number of vectors $x$ that simultaneously minimize $f^\ell(x)$ for all $\ell$ (for any $\ell$).**

3. **Computing the average weight $|x|$ of elements of $\Omega_f$.**

4. **Computing the average Hamming distance $|x - \hat{x}|$ of elements of $\Omega_f$ from a given target $\hat{x}$.**

5. **Computing the number of elements of $\Omega_f$ at a given Hamming distance $k$ from a given target $\hat{x}$.**
Proof:

$$(x_{v_i}, x_{w_i})$$ contributes 1 to the HD

$$\therefore |\Omega_f| = \text{# of } f'\text{-minimizing vectors at HD}$$

k from $\hat{x}$

where $k = |\hat{G}^f_{s,t}| = 3$
Theorem 6 NP-complete to find a fittest seq
\(x\) such that \(k \leq |x| \leq l\)

Proof:

Partially Ordered Knapsack:
\(s(u) = v(u) = \{3, 1, 2\}\)

\[
\sum_{u \in \text{ideal } I} s(u) \leq B \quad \sum_{u \in \text{ideal } I} v(u) \geq K
\]

\(\hat{G}_{s,t}^{f} \cong G_{0} \cong U\)

An ideal of \(\hat{G}_{s,t}^{f}\) corresponds to an ideal of \(U\)

\[
|x| = \sum_{u \in I} |C(u)| = \sum_{u \in I} s(u)
\]

Set \(k = K\) and \(l = B\). \(\blacksquare\)
Applications to Empirical Structures

Relationship between % similarity to native proteins vs. PFAM family size.
Similarity between natural and designed sequences
Open Questions

• Use another folding code (bigger alphabet)

• Improve quality of Fitness Function

Apply model to all PDB structures
Summary of Results

• Enumerate all f-minimizing vectors in $O(n)$ time per vector

• Find diameter in Hamming Distance of all f-minimizing vectors in $O(n)$ time

• Find fittest sequence for $k$ functions or determine if it does not exist

• Compute smallest set of mutations required for adjacency of sequences

• Enumerate sub-optimal sequences by Lawler’s method in $O(n\Delta^2 \log \Delta)$ time per sequence