Biological Overview: Sequence-Structure Asymmetry

- **Horse**
  - Structure: 1DWR
  - Sequence Identity: ~ 85%
  - Protein Source: Sperm Whale

- **Lupinus luteus**
  - Structure: 1LHI
  - Sequence Identity: ~ 20%
  - Protein Source: Sperm Whale
Structures are better conserved than sequences during evolution

- Homology modeling of structures
- Protein design and evolution

<table>
<thead>
<tr>
<th>LH1:</th>
<th>1MBC:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2/3</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1LH1:</td>
<td>1MBC:</td>
</tr>
<tr>
<td>60/61</td>
<td>61/62</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>1LH1:</td>
<td>1MBC:</td>
</tr>
<tr>
<td>118/119</td>
<td>115/116</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Biological Overview:
Sequence-Structure Asymmetry
Biological Overview: Mechanisms of Protein Evolution

The Questions: Sequence Capacity and Flow

- Can we estimate sequence capacity: the number of sequences that are compatible with a given structure?
- Is there migration, or flow, of sequences between structures under point mutations? (May impact protein design and studies of evolution)

Recent paper: Leonid Meyerguz, Jon Kleinberg, and Ron Elber
The network of sequence flow between protein structures
PNAS 2007 104: 11627-11632;
Physical (stability based) network model for sequence capacity of structures & structural flips

- Detailed (whole PDB), efficiently computable and experimentally testable model (the set of PDB structures was argued to be complete)
- Design of protein structures and protein switches
- Zero order model of the evolution of protein sequences & structures (no selection due to function).

Related work on capacity of specific protein models

- Shakhnovich
- Dill
- Wolynes
- Thirumalai
- Levitt
- ...
- So far no global view of capacity (and thermodynamics) of the PDB, no flow.
Is capacity relevant to biology?

- Capacity shows weak correlation with the number of sequences that are found for a particular fold in the NR database (correlation coefficient 0.2)

Experimental tests

- Collaboration with Thomas Magliery on Lambda repressor - 160K mutants
- Protein flips - Bryan lab (flips are well known for RNA)
Protein flips

Patrick A. Alexander, Yanan He, Yihong Chen, John Orban, and Philip N. Bryan “The design and characterization of two proteins with 88% sequence identity but different structure and function” PNAS 2007 104: 11963-11968.

Measuring Protein Fitness: Energy Functions

- Energy functions measure the fitness of sequences to structures.
- If energy is low, then the structure is thermodynamically stable with respect to the probe sequence.

Energy calculation is based on placing amino acids of the $S$ into sites of the target structure $X$.

```
ARNDECQ
```
Approximate Energy Functions

**THOM2** (Meller, Elber)
\[ E(S \rightarrow X) = \sum_{i=1}^{n} c_i^{(X)} (\alpha_i) \]
Ergodic/well mixed model

**TE13** (Toby, Elber)
\[ E(S \rightarrow X) = \sum_{i<j} c(\alpha_i, \alpha_j, r_{ij}^{(X)}) \]
NP complete

Specifying Sequence-Structure Fitness Criteria

Original (Native) Sequence: \( S_{nat}: [HILKMF] \)
Candidate Sequence: \( S: [ARNEDCQ] \)
We say \( S \) is fit for \( X \) if \( E(S \rightarrow X) \leq E(S_{nat} \rightarrow X) \)
Estimating the Sequence Capacity of a Fold

- In general, we would like to compute the function $N(E)$:
  $$N(E) = \left| \left\{ S : E(S \rightarrow X) \leq E \right\} \right|$$

- Specifically, we want to compute $N(E_{\text{nat}})$.
- The number of all possible sequences is $20^n$, for proteins of length $n$. For small proteins, $n \approx 50$.
- Random sampling of sequence space does not work: since $N(E_{\text{nat}}) / 20^n$ can be exponentially small.
- Need a more sophisticated counting method.

Estimating $N(E)$

- Express $N(E) = | \{ S : E(S \rightarrow X) < E_k \} |$ as telescoping ratios/umbrella sampling:
  $$N(E) = N(E_{\text{ref}}) \times \frac{N(E_1)}{N(E_{\text{ref}})} \times \frac{N(E_2)}{N(E_1)} \times \ldots \times \frac{N(E_k)}{N(E_{m})}$$

  - $E_{\text{ref}}$: Pre-selected reference energy
  - $N(E_{\text{ref}})$: Number of sequences below $E_{\text{ref}}$
  - $E_1 \ldots E_m$: Values above ratios are intermediates.
Approximating Successive Ratios

Select site \(i\) u.a.r. from \(S\), and amino acid \(r\) u.a.r. from among the 20 types.

If \(E(S \rightarrow X) + \Delta E < E_k\), accept. Otherwise, reject.

Let \(l_{(k)} = \{ S : E(S \rightarrow X) < E_k \} \) after \(t\) steps.

Then, for sufficiently large \(t\),

\[
\frac{N(E_{k+1})}{N(E_k)} = \frac{l_{(k+1)}}{l_{(k)}}.
\]
Choosing Intermediate $E_k$ Values

Algorithm Summary: $N(E)$

- Given a structure $X$, compute $E_{\text{mean}}$ and $N(E_{\text{mean}})$.
- Pick $E_1 \ldots E_m$ s.t. $E_k > E_{k+1}$ and $(E_k - E_{k+1})$ is decreasing with $k$.
- For $k = 1 \ldots m$, run the Markov chain for $t$ steps. Compute $l_j(k+1) / l_j(k) \approx N(E_{k+1}) / N(E_k)$.

$$N(E) = N(E_{\text{mean}}) \times \frac{N(E_1)}{N(E_{\text{mean}})} \times \frac{N(E_2)}{N(E_1)} \times \ldots \times \frac{N(E)}{N(E_m)}$$
Counting With THOM2: Markov Chain Convergence

- State space is connected: all states communicate via the minimum-energy state
- Mixing time (the Markov chain is ergodic) is polynomial in sequence length.
- Generalizes Morris-Sinclair algorithm (1999) for counting knapsack solutions to arbitrary alphabets

Sequence capacity without competition

- Remain at a particular fold \( X \) and perform counting for this single structure
- Compute \( N(E), \Omega(E)=dN/dE, S(E)=\log(\Omega(E)) \)
- The temperature of sequence selection is defined as: \( T=(dS/dE)^{-1} \)
- Compute for a representative set of PDB structures (~3000 folds)
Weaknesses
(before we even start…)

- No selection due to function
- No domain swaps (only single point mutations)
- Coarse structural models and approximate energy function.
- No structural competition yet (addressed later)

Counting without Competition:
$N(E)$
Counting without competition:
Different folds, same length
(150)

Sequence Capacity and Flow
Coarse description of fold connectivity

- Different temperatures for alternate folds suggests lack of connectivity.

\[ T = (dS/dE)^{-1} \]

Temperature distribution for the potential TE-13
Peaked temperature distribution does not exclude connectivity

Can we suggest a more direct calculation of connectivity between folds?

Counting with Competition

- Fix structure of interest \( X \) as a reference structure.
- Run counting algorithm as described above.
- Keep track of how many sequences have lowest energy in \( X \), and how many “escape” to (achieve lower energy in) competing structures.
- Let \( f_{\text{ret}}(E) \) be the fraction of retained (non-escaping) sequences below energy level \( E \).
- Approximate \( C(E) = N(E) \times f_{\text{ret}}(E) \).
Differentiating Between Competing Folds

- For a fold \( X \), how do we count only sequences that both are both compatible with \( X \) and prefer \( X \) to all other folds?
- Given a structure \( X \) and a set of competing structures \( \mathcal{Y} = \{ Y_1, \ldots, Y_K \} \), we wish to estimate the function \( C(E) \) which gives the size of the set

\[
\{ S \in \Omega : E(S \rightarrow X) < E \land E(S \rightarrow X) < E(S \rightarrow Y_j), \ j = 1 \ldots K \}
\]
Differentiating Between Competing Folds

Counting with Competition

\[ C(E_k) = N(E_k) \times f_{rel}(E_k) \]
Results: Competition Among a Large Library of Protein Folds

- We use a set of 2060 structurally dissimilar protein folds, constituting a representative sample.
- Fix each fold in turn as a reference fold, and run the counting procedure, using the remaining 2059 folds as competitors.
- We are interested in how each fold retains or loses sequences as a function of energy.
- We can model the competition for sequences among folds as a network of sequence flow.

\[ C(E) \text{ and } N(E) \]
Maximal Retention Energy $E^*$

- $E^*$ is the first-encountered energy where $f_{ret}$ is maximum (mostly 1). In general, $E^* \gg E_{min}$.
- Between $E^*$ and $E_{min}$ the protein evolves in structure and sequence spaces.
- Below $E^*$ only the sequence evolves. Native proteins are always found above $E^*$.

$E^*$ and Behavior of $f_{ret}(E)$

- $E_{nat}$
- $E_{mean}$
- $E^*$
- $f_{ret}$
- $1.0$
$E_{\text{nat}} - E^*$ and contact density

$E^*$ and $f_{rel}(E^*)$
Sequence Flow Network

- Nodes are protein structures.
- Edges depend on energy $E$ and a cutoff value $c$.
- There is an edge from $X$ to $Y$ if the fraction of sequences that escape from $X$ into $Y$ at energy level $E$ exceed $c$.
- We are interested in network connectivity at the native energy range.
- Standard cutoff is $c = 1/K$, where $K=2060$ is the number of folds in the dataset.

Example Sequence Flow Network
Proteins sinks are rich in beta sheets

In-Degree Correlations

- Protein length
  - $\rho = 0.630$, P-value $\ll 1e-12$
- $\beta$-Sheet content
  - $\rho = 0.215$, P-value $\ll 1e-12$
  - No correlation between length and $\beta$-content.
- Number of related sequences found by BLAST
  - $\rho = 0.223$, P-value $\ll 1e-12$
Almost ready to speculate

- Modeling kinetics of structural evolution
  (Directed graph and a Master equation at hand -- Many sinks, no origin, folds rich in beta sheet structures are attractors)