Bioinformatics 2 -- Spring 2012

• Lecture 25 -- Protein Design
Protein folding/ protein design

folding

sequence

structure

design
Sequence space maps to structure space

..as many-to-one.
Short history of protein design

Site-directed mutagenesis -- minimization (J. Wells, 1980's-90’s)
Coiled coils, helix bundles (DeGrado, 1980's-90's)
Binary patterning (Hecht, 1990’s)
Extreme protein stabilization (Mayo, 1990's)
Binding pocket design (Hellinga, 2000)
New fold design (Kuhlman & Baker, 2002-4)
Protein-protein interface design (Gray & Baker, 2004)
Open source protein design algorithm EGAD (Pokala, 2005)

Experimental approaches:
- *in vitro* evolution
- phage display

Computational approaches:
- Dead-End elimination
- binary patterning
Proteins can be made super-stable

![Diagram showing comparison between natural and designed protein sequences under denaturing conditions.](image)

Distinct conformational states can be stabilized.

αMβ2 integrin I domain in 2 conformations
2 crystal structures are known. They differ in the highlighted region. Shimaoka _et al_ designed sequences for each form, **open** and **closed**. The two designs were shown to have different physiological properties.

some amazing accomplishments in protein design

Computationally *re-designed* proteins are consistently stable

New folds can be designed that have never been seen before. The designs are accurate (compare red and blue above) and they are highly stable.

New binding sites can be designed

Used to bind arabinose, now it binds serotonin.

Minimizing proteins

(Bing Li et al, Science, 1995)
Binary patterning in proteins

(Kamtekar et al, Science, 1993)
<table>
<thead>
<tr>
<th>Helix 1</th>
<th>Helix 2</th>
<th>Helix 3</th>
<th>Helix 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>MGDLENLLLEKFEQLIKGPDSDKLHNVQKELQEVQGPGSGKWLKNLLNDFEDLINGPRSGNQVQLLKLQMQIKR</td>
<td>.EIIEDELQNLQELME.</td>
<td>.KIQLKIIKVNELMQL.</td>
<td>.DLHNLINKLDDVMQ.</td>
</tr>
<tr>
<td>.DLEKLMDKVNMMIME.</td>
<td>.KFHNHILKELQIMN.</td>
<td>.KLDFHEEMMNKFLK.</td>
<td>.ELHDVHLKLHHLHVMD.</td>
</tr>
<tr>
<td>.EVQRFVFKLEQLLK.</td>
<td>.ELNMKFKEVNLLFK.</td>
<td>.ELERHVIMEDFNMDVE.</td>
<td>.KLEFQRMQHLLOLQ.</td>
</tr>
<tr>
<td>.DMKVEKLQEMQMLD.</td>
<td>.NLQENMELVQDILD.</td>
<td>.QLEEIMMNLENLLQ.</td>
<td>.DILLISECONDSQMNFLQ.</td>
</tr>
<tr>
<td>.ELERFVFKLEQLLK.</td>
<td>.NQLNVLHLLLHNFN.</td>
<td>.NHEVVKMRRKLMNN.</td>
<td>.KVEFQQLQLKHNDN.</td>
</tr>
<tr>
<td>.KNEMLIQELEQLLQ.</td>
<td>.HFFQQLNELHLDIFN.</td>
<td>.DLDFKELLEELLK.</td>
<td>.QVQLQQLKLNIEE.</td>
</tr>
<tr>
<td>.ENMLKLEQHQLKLL.</td>
<td>.QEFKQVMDQKVELE.</td>
<td>.EKLKVMEEFQQLIK.</td>
<td>.ELKHVNNKQQLLIH.</td>
</tr>
<tr>
<td>.NLEKMLEQLOQILQ.</td>
<td>.EVHHLRELQELFE.</td>
<td>.HVNLEKMKKLVKE.</td>
<td>.DVKQNQCLHEH TVQ.</td>
</tr>
<tr>
<td>.KMFLVQQKLHLLK.</td>
<td>.QIKDVQQFQKIMIE.</td>
<td>.EKLKVMEEFQQLIK.</td>
<td>.ELKHVNNKQQLLIH.</td>
</tr>
<tr>
<td>.VEHLEFEEIKIMIE.</td>
<td>.EVQDGFQEQLLHE.</td>
<td>.KMDHMKELQHLE.</td>
<td>.ELKHVNNKQQLLIH.</td>
</tr>
<tr>
<td>.EFHESVKNMQHLK.</td>
<td>.NQIHFLHKIQVLLH.</td>
<td>.ELDKVHHELKNLLE.</td>
<td>.HMQFQKLQFEQVLQ.</td>
</tr>
<tr>
<td>.EMEKFBKEMELE.</td>
<td>.DIEHVKKLMEDFMD.</td>
<td>.KIDKLMKVEHLLN.</td>
<td>.QFENVQVNHELIN.</td>
</tr>
<tr>
<td>.DVEEVEFKMQEVEFH.</td>
<td>.QVQQLKVVNHMMK.</td>
<td>.KLLEELLENNMIIH.</td>
<td>.QQLKQIEFDQMFQ.</td>
</tr>
<tr>
<td>.ELOQQLQVQDELK.</td>
<td>.KRFQQLLMQEHLLN.</td>
<td>.ELSEIMQFQHLLN.</td>
<td>.QFDMKLOQLQELME.</td>
</tr>
<tr>
<td>.DQLHQHIEQQLQK.</td>
<td>.QVQQLKVVNHMMK.</td>
<td>.KLLEELLENNMIIH.</td>
<td>.QQLKQIEFDQMFQ.</td>
</tr>
<tr>
<td>.EIQQQQVLQQQI.</td>
<td>.HVVQIIMEQMNMLQ.</td>
<td>.QMDVQLQEQMNMME.</td>
<td>.HVNVFQMNQLIN.</td>
</tr>
<tr>
<td>.IDQVVQVQEHKLVD.</td>
<td>.HLKNSQMQQIQIIQ.</td>
<td>.NMENLLEQLEEFK.</td>
<td>.DLQKIVHDFDKMLN.</td>
</tr>
<tr>
<td>.NLDELFEEKLQMLE.</td>
<td>.HIDKLMENLKQMLQ.</td>
<td>.ELEFLKIEDLLK.</td>
<td>.KLQILQIELEDNFN.</td>
</tr>
<tr>
<td>.EVENILKQELVBE.</td>
<td>.NLKDQINLKQLEE.</td>
<td>.ELDFKQLKELLLK.</td>
<td>.QVQIQVIIHHIQFLQ.</td>
</tr>
<tr>
<td>.RMNDIHLDDLQME.</td>
<td>.QINEFVHQMEMBE.</td>
<td>.EIQIIIDEDLQLE.</td>
<td>.QEHLIQKFFHLE.</td>
</tr>
<tr>
<td>.EFQEMKLEDLMFH.</td>
<td>.ELEQFHEIQEFBN.</td>
<td>.KLDQQLEEEVDILT.</td>
<td>.QLHELLQDMMHLVQ.</td>
</tr>
<tr>
<td>.DLEDDLEEMQLNQL.</td>
<td>.HMLQQDLDFQFVE.</td>
<td>.DMQKLLNDVMKELN.</td>
<td>.DFQMNQHQLHNVLD.</td>
</tr>
<tr>
<td>.KLKLMHQFOQQLLQ.</td>
<td>.DIKHLNHMEMKLHN.</td>
<td>.ELHNPLHNLHEHLL.</td>
<td>.NVQKLQVDQVLHFN.</td>
</tr>
<tr>
<td>.KLDNDDLEDQME.</td>
<td>.HLQNVEIHDQMQ.</td>
<td>.KLQEMMKEFFQVQL.</td>
<td>.NKEYIFHHLEELVH.</td>
</tr>
<tr>
<td>.MGHLEEILNEMQEMGLDPDSQVVKQILNELIQMLEGPSGGHQMNIFKMNHLKFLQGPRSGQVEQIQFEKLHKFFHR</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Polar residues**

**Nonpolar residues**
Computational protein design using Dead-End Elimination

- Select positions for mutating.
- Select allowed amino acids at those positions.
- For the selected amino acids, try all sidechain orientations (rotamers).

Chose the set of rotamers that gives the lowest “energy”
Sidechain Rotamers

Sidechain conformations fall into three classes called rotational isomers, or **rotamers**.

A random sampling of Phenylalanine sidechains, w/ backbone superimposed
Sidechain rotamers

1-4 interactions differ greatly in energy depending on the moieties involved.

"m" -60° gauche

"t" 180° anti/trans

"p" +60° gauche
Rotamer stability is dependent on the backbone $\phi\psi$ angles

W sidechain is shown here lying over Thr backbone

<table>
<thead>
<tr>
<th>Rotamers of W*: $\phi$</th>
<th>$\psi$</th>
<th>$P_{\phi=-140,\psi=160}$</th>
<th>$P_{\phi=-60,\psi=-40}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>p-90</td>
<td>+60</td>
<td>-90</td>
<td>0.372</td>
</tr>
<tr>
<td>p90</td>
<td>+60</td>
<td>+90</td>
<td>0.238</td>
</tr>
<tr>
<td>t-105</td>
<td>180</td>
<td>-105</td>
<td>0.033</td>
</tr>
<tr>
<td>t90</td>
<td>180</td>
<td>90</td>
<td>0.021</td>
</tr>
<tr>
<td>m0</td>
<td>-65</td>
<td>5</td>
<td>0.038</td>
</tr>
<tr>
<td>m95</td>
<td>-65</td>
<td>95</td>
<td>0.183</td>
</tr>
</tbody>
</table>
Rotamer Libraries

Rotamer libraries have been compiled by clustering the sidechains of each amino acid over the whole database. Each cluster is a representative conformation (or rotamer), and is represented in the library by the best sidechain angles (chi angles), the "centroid" angles, for that cluster.

Two commonly used rotamer libraries:

* Jane & David Richardson:  http://kinemage.biochem.duke.edu/databases/rotamer.php

*rotamers of W on the previous page are from the Richardson library.
Given the sequence and only the backbone atom coordinates, accurately model the *positions of the sidechains*.

**fine lines** = true structure  
**thick lines** = sidechain predictions using the method of Desmet et al.

Theoretical complexity of sequence design

Estimated number of sidechain rotamers: $R=193$

Typical small protein length: $L=100$ residues

Sequence complexity: $20^{100} = 1.3 \times 10^{130}$

Rotamer complexity: $193^{100} = 3.6 \times 10^{228}$

*Complexity of DEE algorithm: $O(R^2L^2) = 3.6 \times 10^8$*
Dead end elimination theorem

• Each residue is numbered \((i\) or \(j\)) and each residue has a set of rotamers \((r, s\) or \(t\)). So, the notation \(i_r\) means "choose rotamer \(r\) for position \(i\)."

• The total energy is the sum of the three components:

\[
E_{\text{global}} = E_{\text{template}} + \sum_i E(i_r) + \sum_i \sum_j E(i_r, j_s)
\]

where \(r\) and \(s\) are any choice of rotamers.

NOTE: \(E_{\text{global}} \geq E_{\text{GMEC}}\) for any choice of rotamers.
Dead end elimination theorem

• If $i_g$ is in the GMEC and $i_t$ is not, then we can separate the terms that contain $i_g$ or $i_t$ and re-write the inequality.

\[
E_{\text{GMEC}} = E_{\text{template}} + E(i_g) + \sum_j E(i_g,j_g) + \sum_j E(j_g) + \sum_j \sum_k E(j_g,k_g)
\]

...is less than...

\[
E_{\text{notGMEC}} = E_{\text{template}} + E(i_t) + \sum_j E(i_t,j_g) + \sum_j E(j_g) + \sum_j \sum_k E(j_g,k_g)
\]

Canceling all terms in black, we get:

\[
E(i_r) + \sum_j E(i_r,j_s) > E(i_g) + \sum_j E(i_g,j_s)
\]

So, if we find two rotamers $i_r$ and $i_t$, and:

\[
E(i_r) + \sum_j \min_s E(i_r,j_s) > E(i_t) + \sum_j \max_s E(i_t,j_s)
\]

Then $i_r$ cannot possibly be in the GMEC.
Dead end elimination theorem

\[ E(i_r) + \sum_j \min_s E(i_rj_s) > E(i_t) + \sum_j \max_s E(i_tj_s) \]

DEE theorem can be translated into plain English as follows:

If the "worst case scenario" for \( t \) is better than the "best case scenario" for \( r \), then you always choose \( t \).
Find two columns (rotamers) within the same residue, where one is always better than the other. Eliminate the rotamer that can always be beat. (repeat until only 1 rotamer per residue)
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>a</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>c</td>
<td>c</td>
<td>c</td>
<td></td>
</tr>
</tbody>
</table>

\[
E(r_{1},r_{2})
\]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>3</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>5</td>
<td>5</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>-2</td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>5</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>12</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

\[
E(r_{1})
\]

\[
E(r_{2})
\]
DEE algorithm

\[ E_{\text{global}} = E_{\text{template}} + \sum_i E(i_r) + \sum_i \sum_j E(i_r, j_s) \]

Find two columns (rotamers) within the same residue, where one is always better than the other. Eliminate the rotamer that can always be beat. (repeat until only 1 rotamer per residue)
“Rotamers” within the DEE framework can have different atoms, i.e. they can be different amino acids. Using DEE, we choose the best set of rotamers. Now we have the sequence of the lowest energy structure. In the example, we have D or L at position 3.
Recent success of sequence design

Re-designing a binding site

(a) Diagram showing a protein structure with a ribose molecule labeled. The diagram includes sequences of amino acids and chemical structures for TNT, L-Lactate, and Serotonin.

(b) Chemical structures of TNT, L-Lactate, and Serotonin.
Each alternative ligand position is another “rotamer”.
An appropriate binding site was found

The native ligand (arabinose) is approximately the same size as the targeted ligand (seratonin).

A space was carved out for the ligand

All sidechains in the binding site were truncated to alanines, and a space was defined (yellow) for the new ligand. Lots of possible ligand orientations were made. **Ligand orientations were treated like rotamers in DEE!**

Sidechains were chosen using DEE

The most critical component of the energy function was *hydrogen bonding* (dotted lines). Every donor/acceptor should be satisfied.

H-bonding is key
New directions for protein design

• Docking + design
• Receptor design -- biosensors
• Designing kinetic stability -- re-wiring
Programmable Fluorescent biosensors

A + B = C
Green Fluorescent Protein

11-standed beta-barrel surrounding fluorescent chromophore
GFP biosensor showing bound target peptide
GFP is circularly permuted using a C-N linker peptide (red arrow), and the majority of β-strand 7 is removed (black arrow). Coordinates were modeled from PDB 1EMA using MOE.
Proof of biosensor designability

Target peptide (green) bound to designed biosensor with 4 mutations (yellow)

Colonies of bacteria making designed GFP biosensors. At least 20 of the colonies are glowing.
**Target specificity**: Biosensor t7SP incubated with target peptide s7. Fluorescence grows in proportional to s7 concentration.
Multi-analyte test strips
Engineering kinetic stability by rewiring proteins

GFP

Re-wired GFP