Accurate prediction for atomic-level protein design and its application in diversifying the near-optimal sequence space

Pablo Gainza
CPS 296: Topics in Computational Structural Biology
Department of Computer Science
Duke University
Outline

1) Problem definition
2) Formulation as an inference problem
3) Graphical Models
4) tBMMF algorithm
5) Results
6) Conclusions
1. Problem Definition

Protein design algorithm

\[ G \text{MEC} \in S \]
1. Problem Definition (2)

\[ r \rightarrow \text{Rotamer assignment (RA)} \]

\[ r_i \rightarrow \text{Rotamer at position } i \text{ for RA } r \]
1. Problem Definition (3)

\[ E_i(r_i) \rightarrow \text{Energy between rotamer } r_i \text{ and fixed backbone} \]

\[ E_{ij}(r_i, r_j) \rightarrow \text{Energy between rotamers } r_i \text{ and } r_j \]

\[ E(r) \rightarrow \text{Energy of rotamer assignment } r \]

\[ E(r) = \sum_i E_i(r_i) + \sum_{i,j} E_{ij}(r_i, r_j) \]
1. Problem Definition (4)

\[ T(k) \rightarrow \text{returns amino acid type of rotamer } k \]

\[ T(r) \rightarrow \text{returns sequence of rotamer assignment } r \]

\[ T(r_1) = \text{hexagon} \]

\[ T(r_2) = \text{cross} \]

\[ T(r) = \text{hexagon, cross} \]
1. Problem Definition (5)

- **Rotamer Library**
- **Positions to design & allowed Rotamers/Amino Acids**
- **Energy f(x)**
- **Protein structure**
- **Protein design algorithm**

\[
S^* = T(\arg \min_r E(r))
\]

\[
\text{GMEC} \in S
\]
1. Problem Definition (6)

Related Work

Exact Methods

Global Minimum Energy Conformation

Low energy conformation

Probabilistic Methods

DEE / A*
BroMAP
BWM

SCMF
MCSA
1. Problem Definition (7)

Rotamer Library

Energy $f(x)$

Protein structure

Positions to design & allowed Rotamers/Amino Acids

Model → Inaccurate!
1. Problem Definition (8)

\[ S^* = T(\arg\min_r E(r)) \]
1. Problem Definition (9)

- Too stable
- Not fold to target
- Low energy conformation
- Low binding specificity
Solution: Find a set of low energy sequences

Provable Methods
- DEE/A*

Ordered set of gap-free low energy conformations, including GMEC

Set of low energy conformations

Probabilistic Methods
- tBMMF
Problem Definition: Summary

• Protein design algorithms search for the **sequence** with the Global Minimum Energy Conformation (**GMEC**).

• Our model is **inaccurate**: more than one low energy sequence is desirable.

• Fromer et al. Propose **tBMMF** to generate a **set** of low energy sequences.
2. Our problem as an inference problem

**Probabilistic factor** for self-interactions

\[ \psi_i(r_i) = e^{\frac{-E_i(r_i)}{T}} \]

**Probabilistic factor** for pairwise interactions

\[ \psi_{ij}(r_i, r_j) = e^{\frac{-E_{ij}(r_i, r_j)}{T}} \]
2. Inference problem (2)

\[ Z = \sum_r e^{\frac{E(r)}{T}} \]

Probability \textbf{distribution} for rotamer assignment \( r \)

\[ P(r_1, \ldots, r_N) = \frac{1}{Z} \prod_i \psi_i(r_i) \prod_{i,j} \psi_{ij}(r_i, r_j) = \frac{1}{Z} e^{\frac{-E(r)}{T}} \]
2. Inference problem (3)

Minimization goal (from \textbf{definition})

\[ S^* = T(\arg \min_r E(r)) \]

Minimization goal for a graphical model problem

\[ S^* = T(\arg \max_r Pr(r)) \]
2. Inference problem (4)

\[ E_{ij}(r_1, r_2) \]

<table>
<thead>
<tr>
<th>Position #1</th>
<th>Position #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>-2</td>
</tr>
</tbody>
</table>

\[
E_i(r_1) \quad E_i'(r_2)
\]

<table>
<thead>
<tr>
<th>r_0</th>
<th>r_0'</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-4</td>
<td>-5</td>
</tr>
<tr>
<td>-2</td>
<td>-3</td>
</tr>
</tbody>
</table>

Example: Inference problem

Allowed

\[ E(r') =? \]

\[ E(r'') =? \]

What is our GMEC??
2. Inference problem (5)

\[ E_{ij}(r_1, r_2) \]

Position #1

\[
\begin{array}{c|c|c}
\text{Position #2} & \text{ Allowed} \\
-4 & 1 & 2 \\
-2 & 1 & 2 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c|c}
\text{ } & \text{ } & \text{ } & \text{ } & \text{ } \\
E_i(r_1) & E_i(r_2) & E_i(r_0) & E_i(r_0) & E_i(r_0) \\
-1 & -5 & -3 & -1 & -2 \\
\end{array}
\]

\[
E(r') = (-1 + -2) + (-5 + -2) = -10
\]

\[
E(r'') = (-1 + -4) + (-3 + -4) = -12
\]

\[ r'' \text{ is our GMEC} \]
2. Inference problem (6)

\[ E_{ij}(r_1, r_2) \]

Position #1

<table>
<thead>
<tr>
<th></th>
<th>Allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>1</td>
</tr>
<tr>
<td>-2</td>
<td>2</td>
</tr>
</tbody>
</table>

Position #2

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
</tr>
</tbody>
</table>

\[ E_i(r_1) \quad E_i(r_2) \]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-5</td>
</tr>
<tr>
<td></td>
<td>-3</td>
</tr>
</tbody>
</table>

\[ T = 1 \text{ (for our example)} \]

\[
\begin{align*}
\psi_i(r'_1) &= e \frac{-E_i(r'_1)}{T} = e \\
\psi_i(r'_2) &= e \frac{-E_i(r'_2)}{T} = e^5 \\
\psi_i(r''') &= e \frac{-E_i(r'''}{T} = e \\
\psi_i(r'') &= e \frac{-E_i(r'')}{T} = e^3
\end{align*}
\]
2. Inference problem (7)

\[ E_{ij}(r_1, r_2) \]

<table>
<thead>
<tr>
<th>Position #2</th>
<th>Allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>1</td>
</tr>
<tr>
<td>-2</td>
<td>2</td>
</tr>
</tbody>
</table>

\[ E_i(r_1) \quad E_i(r_2) \]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-5</td>
</tr>
<tr>
<td></td>
<td>-3</td>
</tr>
</tbody>
</table>

\[ T = 1 \] (for our example)

\[
\psi_{ij}(r_1', r_2') = e^{\frac{-E_{ij}(r_1', r_2')}{T}} = e^2
\]

\[
\psi_{ij}(r_1'', r_2'') = e^{\frac{-E_{ij}(r_1'', r_2'')}{T}} = e^4
\]

\[
Z = \sum_r e^{\frac{E(r)}{T}} = e^{10} + e^{12}
\]
2. Inference problem (8)

\[ E_{ij}(r_1, r_2) \]

\[ \begin{array}{c|cc}
\text{Position \#1} & -4 & -2 \\
\text{Position \#2} & -1 & -5 & -3 \\
\end{array} \]

\[ P(r'_1, r'_2) = \frac{1}{Z} \prod_i \psi_i(r'_i) \prod_{i,j} \psi_{ij}(r'_i, r'_j) \]

\[ = \frac{e^{10}}{e^{10} + e^{12}} \]

\[ T = 1 \text{ (for our example)} \]
2. Inference problem (9)

\[ E_{ij}(r_1, r_2) \]

Position #1

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td>-2</td>
</tr>
</tbody>
</table>

Position #2

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-5</td>
</tr>
<tr>
<td>-3</td>
<td></td>
</tr>
</tbody>
</table>

\[ E_i(r_1) \quad E_i(r_2) \]

\[ P(r''_1, r''_2) = \frac{1}{Z} \prod_i \psi_i(r_i') \prod_{i,j} \psi_{ij}(r''_i, r''_j) \]

\[ = \frac{e^{12}}{e^{10} + e^{12}} \]

\[ T = 1 \text{ (for our example)} \]
2. Inference problem (10)

\[ E_{ij}(r_1, r_2) \]

<table>
<thead>
<tr>
<th>Position #1</th>
<th></th>
<th>Position #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r_1 )</td>
<td>( r_2 )</td>
<td>( r_0 )</td>
</tr>
<tr>
<td>-4</td>
<td>-2</td>
<td>-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( E_i(r_1) )</th>
<th>( E_i(r_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-5</td>
</tr>
<tr>
<td>-3</td>
<td></td>
</tr>
</tbody>
</table>

Allowed

\( r' \)

\( r'' \)

\( S^* = T(\arg \max_r Pr(r)) \)

\( T = 1 \) (for our example)
Minimization goal (from definition)

\[ S^* = T(\arg \min_r E(r)) \]

Minimization goal for a graphical model problem

\[ S^* = T(\arg \max_r Pr(r)) \]

We still have a non-polynomial problem!

But formulated as an inference problem

Probabilistic methods
Summary: Inference problem

• We model our problem as an *inference* problem.

• We can use *probabilistic* methods to solve it.
3. Graphical models for protein design and belief propagation (BP)

1. Model each design position as a random variable

2. Build interaction graph that shows conditional independence between variables


SspB dimer interface: Inter-monomeric interactions (Cα)
Example: Belief propagation

node in the graphical model: interacting residue in the structure.

node in the graphical model: random variable
Example: Belief propagation

edge: energy interaction between two residues.

edge: causal relationship between two nodes

If two residues are distant from each other, no edge between them.
Every random variable can be in one of several states: allowable rotamers for that position.
The energy of each state depends on:
- its **singleton** energy
- its **pairwise** energies
- the energies of the states of its parents
Example: Belief propagation

Belief propagation: each node tells its neighbors nodes what it believes their state should be.

A message is sent from node i to node j.

The message is a vector where # of dimensions: allowed states/rotamers in recipient.
Example: Belief propagation

Who sends the first message?
Who sends the first message?

In a tree: the leaves
- Belief propagation is proven to be correct in a tree!
Who sends the first message?

In a graph with cycles:
- Set initial values
- Send in parallel

No guarantees can be made! There might not be any convergence
3. Graphical Models/BP (10)

Example: Belief propagation

We iterate from there.

\[
\begin{align*}
\hat{m}_{1 \rightarrow 3}(r'_{3}) &= 1 \\
\hat{m}_{2 \rightarrow 3}(r''_{3}) &= 1
\end{align*}
\]

\[
m_{r_{1} \rightarrow r}(r'_{1}) = 1
\]

\[
m_{r_{3} \rightarrow r}(r'_{3}) = 1
\]

\[
m_{r_{2} \rightarrow r}(r'_{2}) = 1
\]

\[
m_{r_{3} \rightarrow 1}(r'_{3}) = 1
\]

\[
m_{1 \rightarrow r}(r'_{1}) = 1
\]

\[
m_{2 \rightarrow 3}(r''_{3}) = 1
\]

\[
m_{1 \rightarrow 3}(r'_{3}) = 1
\]
3. Graphical Models/BP (11)

Example: Belief propagation

Node 3 receives messages from nodes 1 and 2

\[
\begin{align*}
m_{\gamma \rightarrow \nu}(r_{\nu}) &= 1 \\
m_{\gamma \rightarrow \nu}(r_{\nu}') &= 1
\end{align*}
\]
What message does node 3 send to node 1 on the next iteration?

$m_{3 \rightarrow 1}(r'_1) = ?$
Belief propagation: message passing

\[ N(i) \rightarrow \text{Neighbors of variable } i \]

Message that gets sent on each iteration

\[
m_{i\rightarrow j}(r_j) = \max_{r_i} \left( e^{\frac{-E_i(r_i) - E_{ij}(r_i, r_j)}{t}} \prod_{k \in N(i) \setminus j} m_{k\rightarrow i}(r_i) \right)
\]
**Example: Belief propagation**

### Pairwise energies

\[ E_{ij}(r_1, r_2) \]

Position #1

\[ E_{ij}(r_2, r_3) \]

Position #3

\[ E_{ij}(r_1, r_3) \]

Position #3

### Singleton energies

<table>
<thead>
<tr>
<th></th>
<th>( E_i(r_1) )</th>
<th>( E_i(r_2) )</th>
<th>( E_i(r_3) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1</td>
<td>-2</td>
<td>-6 ( r_3' )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-2 ( r_3'' )</td>
</tr>
</tbody>
</table>

### Iteration 0:

\[
m_{3 \rightarrow 1}(r_1') = \max_{r_3} \left( e^{\frac{-E_i(r_3) - E_{ij}(r_3', r_1)}{t}} m_{2 \rightarrow 3}(r_3'), e^{\frac{-E_i(r_3) - E_{ij}(r_3'', r_1)}{t}} m_{2 \rightarrow 3}(r_3'') \right) = ?
\]
3. Graphical Models/BP (15)

Example: Belief propagation

Once it converges we can compute the belief each node has about **itself**

Belief about one's state:
Multiply all incoming **messages by singleton** energy
Belief propagation: Max-marginals

Belief about each rotamer

\[ MM_i(r_i) = e^{\frac{-E_i(r_i)}{t}} \prod_{k \in N(i)} m_{k\rightarrow i}(r_i) \]

“Most likely” rotamer for position i

\[ r_i^* = \arg \max_{r_i \in \text{Rot}_i} \max_{r' : r'_i = r_i} \Pr_i(r'_i) \]
3. Graphical Models/BP (17)

Fromer M, Yanover, C. Proteins (2008)

Fromer M, Yanover, C. Proteins (2008)
3. Graphical Models/BP (18)

Fromer M, Yanover, C. Proteins (2008)

Fromer M, Yanover, C. Proteins (2008)
3. Graphical Models: Summary

- Formulate as an inference problem
- Model our design problem as a graphical model
- Establish edges between interacting residues
- Use Belief Propagation to find the beliefs for each position
4. tBMMF: type specific BMMF

• Paper's main contribution
• Builds on previous work by C. Yanover (2004)
• Uses **Belief propagation** to find lowest energy sequence and **constrains** space to find subsequent sequences
TBMMF (simplification)

1. Find the lowest energy sequence using BP

2. Find the next lowest energy sequence while excluding amino acids from the previous one

3. Partition into two subspaces using constraints according to the next lowest energy sequence
4. tBMMF (3)

Example: tBMMF (1)

### A

<table>
<thead>
<tr>
<th>aa</th>
<th>rot.</th>
<th>Position #1</th>
<th>( G_1 )</th>
<th>( G_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( g_{11} )</td>
<td>( g_{12} )</td>
<td>( g_{21} )</td>
</tr>
<tr>
<td>( H_1 )</td>
<td>( h_{11} )</td>
<td>-15</td>
<td>-11</td>
<td>-6</td>
</tr>
<tr>
<td></td>
<td>( h_{12} )</td>
<td>-14</td>
<td>-10</td>
<td>-7</td>
</tr>
<tr>
<td>( H_2 )</td>
<td>( h_{21} )</td>
<td>-8</td>
<td>-9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>( h_{22} )</td>
<td>-12</td>
<td>-13</td>
<td>-4</td>
</tr>
</tbody>
</table>

### B

<table>
<thead>
<tr>
<th>( r )</th>
<th>( E(r) )</th>
<th>( T(r) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (g_{11}, h_{11}) )</td>
<td>-15</td>
<td>( (G_1, H_1) )</td>
</tr>
<tr>
<td>( (g_{11}, h_{12}) )</td>
<td>-14</td>
<td>( (G_1, H_1) )</td>
</tr>
<tr>
<td>( (g_{12}, h_{22}) )</td>
<td>-13</td>
<td>( (G_1, H_2) )</td>
</tr>
<tr>
<td>( (g_{11}, h_{22}) )</td>
<td>-12</td>
<td>( (G_1, H_2) )</td>
</tr>
<tr>
<td>( (g_{12}, h_{11}) )</td>
<td>-11</td>
<td>( (G_1, H_1) )</td>
</tr>
<tr>
<td>( (g_{12}, h_{12}) )</td>
<td>-10</td>
<td>( (G_1, H_1) )</td>
</tr>
<tr>
<td>( (g_{12}, h_{21}) )</td>
<td>-9</td>
<td>( (G_1, H_2) )</td>
</tr>
<tr>
<td>( (g_{11}, h_{21}) )</td>
<td>-8</td>
<td>( (G_1, H_2) )</td>
</tr>
<tr>
<td>( (g_{21}, h_{12}) )</td>
<td>-7</td>
<td>( (G_2, H_1) )</td>
</tr>
<tr>
<td>( (g_{21}, h_{11}) )</td>
<td>-6</td>
<td>( (G_2, H_1) )</td>
</tr>
<tr>
<td>( (g_{22}, h_{21}) )</td>
<td>-5</td>
<td>( (G_2, H_2) )</td>
</tr>
<tr>
<td>( (g_{21}, h_{22}) )</td>
<td>-4</td>
<td>( (G_2, H_2) )</td>
</tr>
<tr>
<td>( (g_{22}, h_{11}) )</td>
<td>-3</td>
<td>( (G_2, H_1) )</td>
</tr>
<tr>
<td>( (g_{22}, h_{12}) )</td>
<td>-2</td>
<td>( (G_2, H_1) )</td>
</tr>
<tr>
<td>( (g_{22}, h_{22}) )</td>
<td>-1</td>
<td>( (G_2, H_2) )</td>
</tr>
<tr>
<td>( (g_{21}, h_{21}) )</td>
<td>0</td>
<td>( (G_2, H_2) )</td>
</tr>
</tbody>
</table>

Fromer M, Yanover, C. Proteins (2008)
4. tBMMF (4)

Example: tBMMF (2)

Fromer M, Yanover, C. Proteins (2008)
Results

prion
1H4M  
(a+b)

SspB
1OU9  
all beta

hGH-hGHR
3HHR  
all alpha

CaM-smMLCK
1CDL  
all alpha

CaM-skMLCK
2BBM  
all alpha

Top7
1QYS  
(a+b)

Fromer M, Yanover, C. Proteins (2008)
# Results (2)

<table>
<thead>
<tr>
<th></th>
<th>Num. Positions (Chains&lt;sup&gt;a&lt;/sup&gt;)</th>
<th>Search Space Cardinality (log&lt;sub&gt;10&lt;/sub&gt;)</th>
<th>Rotamer Library</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Design</td>
<td>Shell&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Sequence</td>
</tr>
<tr>
<td><strong>SMALL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>prion</td>
<td>7 (A)</td>
<td>7 (B)</td>
<td>8.95</td>
</tr>
<tr>
<td>SspB</td>
<td>8 (A,C)</td>
<td>0</td>
<td>10.23</td>
</tr>
<tr>
<td><strong>MEDIUM</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hGH-hGHR 1</td>
<td>6 (A)</td>
<td>135 (A,B)</td>
<td>7.67</td>
</tr>
<tr>
<td>hGH-hGHR 2</td>
<td>6 (A)</td>
<td>135 (A,B)</td>
<td>7.67</td>
</tr>
<tr>
<td>hGH-hGHR 3</td>
<td>5 (A)</td>
<td>136 (A,B)</td>
<td>6.39</td>
</tr>
<tr>
<td>hGH-hGHR 4</td>
<td>6 (A)</td>
<td>135 (A,B)</td>
<td>7.67</td>
</tr>
<tr>
<td>hGH-hGHR 5</td>
<td>6 (A)</td>
<td>135 (A,B)</td>
<td>7.67</td>
</tr>
<tr>
<td>hGH-hGHR 6</td>
<td>6 (A)</td>
<td>135 (A,B)</td>
<td>7.67</td>
</tr>
<tr>
<td><strong>LARGE 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CaM-smMLCK</td>
<td>24 (A)</td>
<td>19 (B)</td>
<td>30.69</td>
</tr>
<tr>
<td>CaM-skMLCK</td>
<td>24 (A)</td>
<td>19 (B)</td>
<td>30.69</td>
</tr>
<tr>
<td><strong>LARGE 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hGH-hGHR</td>
<td>35 (A)</td>
<td>106 (A,B)</td>
<td>44.76</td>
</tr>
<tr>
<td>Top7</td>
<td>92 (A)</td>
<td>0</td>
<td>117.65</td>
</tr>
</tbody>
</table>

<sup>a</sup>Peptide chains to which the corresponding positions belong, labeled arbitrarily.

<sup>b</sup>Non-designed, conformationally varying positions.

<sup>c</sup>Rotamer space cardinality after application of type-dependent Goldstein DEE.

<sup>d</sup>Full: all rotamers read from library; Limited: highest probability rotamers read.

<sup>e</sup>Side-chain angles around which additional rotamers were super-sampled from library rotamers.

*Fromer M, Yanover, C. Proteins (2008)*
Results(3)

• Algorithms tried:
  – DEE / A* (Goldstein, 1-split, 2-split, Magic Bullet)
  – tBMMF
  – Ros: Rosetta
  – SA: Simulated annealing over sequence space
Results (4): Assessment results

Fromer M, Yanover, C. Proteins (2008)
Results (5)

Fromer M, Yanover, C. Proteins (2008)
Table I. Assessment and Analysis of the Algorithms Tested

<table>
<thead>
<tr>
<th></th>
<th>tBMMF</th>
<th>Ros</th>
<th>SA</th>
<th>A*</th>
<th>(A* Rotamer Space)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top</td>
<td>Time</td>
<td>Top</td>
<td>Time</td>
<td>Top$^a$</td>
</tr>
<tr>
<td>Small prion</td>
<td>100%</td>
<td>58.9 m</td>
<td>86%</td>
<td>9.3 h</td>
<td>100%</td>
</tr>
<tr>
<td>SspB</td>
<td>100%</td>
<td>11 h</td>
<td>100%</td>
<td>11.4 h</td>
<td>97%</td>
</tr>
<tr>
<td>Medium hGH-hGHR 1</td>
<td>88%</td>
<td>13.4 h</td>
<td>30%</td>
<td>2.1 d</td>
<td>2%</td>
</tr>
<tr>
<td>hGH-hGHR 2</td>
<td>60%</td>
<td>7.6 h</td>
<td>50%</td>
<td>2 d</td>
<td>0%</td>
</tr>
<tr>
<td>hGH-hGHR 3</td>
<td>100%</td>
<td>4.1 h</td>
<td>73%</td>
<td>1.7 d</td>
<td>0%</td>
</tr>
<tr>
<td>hGH-hGHR 4</td>
<td>100%</td>
<td>8.5 h</td>
<td>22%</td>
<td>2.1 d</td>
<td>0%</td>
</tr>
<tr>
<td>hGH-hGHR 5</td>
<td>100%</td>
<td>2.9 h</td>
<td>27%</td>
<td>2 d</td>
<td>0%</td>
</tr>
<tr>
<td>hGH-hGHR 6</td>
<td>100%</td>
<td>8.5 h</td>
<td>42%</td>
<td>2.2 d</td>
<td>0%</td>
</tr>
<tr>
<td>Large 1 CaM-smMLCK</td>
<td>73%</td>
<td>10.6 h</td>
<td>18%</td>
<td>18 h</td>
<td>23%</td>
</tr>
<tr>
<td>CaM-skMLCK</td>
<td>100%</td>
<td>2 h</td>
<td>0%</td>
<td>10.7 h</td>
<td>0%</td>
</tr>
<tr>
<td>Large 2 hGH-hGHR</td>
<td>100%</td>
<td>17.6 h</td>
<td>0%</td>
<td>2 d</td>
<td>0%</td>
</tr>
<tr>
<td>Top7</td>
<td>60%</td>
<td>7 h</td>
<td>3%</td>
<td>15.4 h</td>
<td>0%</td>
</tr>
</tbody>
</table>
## Results (6)

Table I. Assessment and Analysis of the Algorithms Tested

<table>
<thead>
<tr>
<th></th>
<th>tBMMF</th>
<th>Ros</th>
<th>SA</th>
<th>A*</th>
<th>(A* Rotamer Space)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top</td>
<td>Time</td>
<td>Top</td>
<td>Time</td>
<td>Top</td>
</tr>
<tr>
<td>Small prion</td>
<td>100%</td>
<td>58.9 m</td>
<td>86%</td>
<td>9.3 h</td>
<td>100%</td>
</tr>
<tr>
<td>SspB</td>
<td>100%</td>
<td>11 h</td>
<td>100%</td>
<td>11.4 h</td>
<td>97%</td>
</tr>
<tr>
<td>Medium hGH-hGHR 1</td>
<td>88%</td>
<td>13.4 h</td>
<td>30%</td>
<td>2.1 d</td>
<td>2%</td>
</tr>
<tr>
<td>hGH-hGHR 2</td>
<td>60%</td>
<td>7.6 h</td>
<td>50%</td>
<td>2 d</td>
<td>0%</td>
</tr>
<tr>
<td>hGH-hGHR 3</td>
<td>100%</td>
<td>4.1 h</td>
<td>73%</td>
<td>1.7 d</td>
<td>0%</td>
</tr>
<tr>
<td>hGH-hGHR 4</td>
<td>100%</td>
<td>8.5 h</td>
<td>22%</td>
<td>2.1 d</td>
<td>0%</td>
</tr>
<tr>
<td>hGH-hGHR 5</td>
<td>100%</td>
<td>2.9 h</td>
<td>27%</td>
<td>2 d</td>
<td>0%</td>
</tr>
<tr>
<td>hGH-hGHR 6</td>
<td>100%</td>
<td>8.5 h</td>
<td>42%</td>
<td>2.2 d</td>
<td>0%</td>
</tr>
<tr>
<td>Large 1 CaM-smMLCK</td>
<td>73%</td>
<td>10.6 h</td>
<td>18%</td>
<td>18 h</td>
<td>23%</td>
</tr>
<tr>
<td>CaM-skMLCK</td>
<td>100%</td>
<td>2 h</td>
<td>0%</td>
<td>10.7 h</td>
<td>0%</td>
</tr>
<tr>
<td>Large 2 hGH-hGHR</td>
<td>100%</td>
<td>17.6 h</td>
<td>0%</td>
<td>2 d</td>
<td>0%</td>
</tr>
<tr>
<td>Top7</td>
<td>69%</td>
<td>7.1 h</td>
<td>31%</td>
<td>1.5 d</td>
<td>0%</td>
</tr>
</tbody>
</table>

Fromer M, Yanover, C. Proteins (2008)
Results (7)

- DEE/A* was **not feasible** for any case except the prion
- SspB: A* could only output one sequence
- DEE also did not finish after 12 days
- BD/K* did not finish after 12 days
Results (8)

- Predicted sequences where **highly similar** between themselves. (high sequence identity)
- Very **different** from **wild type** sequence
- Solution: *grouped tBMMF*: apply constraints to whole groups of amino acids – proof of concept only
Conclusions

• Fast and accurate algorithm
• Outperforms all other algorithms:
  – A* is not feasible
  – Better accuracy than other probabilistic algorithms
Conclusions (2)

- tBMMF produces a large set of very similar low energy results.
- This might be due to the many inaccuracies in the model
- Grouped tBMMF can produce a diverse set of low energy sequences
Conclusions (3)

- The results lack experimental data for validation.
Related Work: (Fromer et al. 2008)

- Fromer F, Yanover C. *A computational framework to empower probabilistic protein design*. ISMB 2008

- **Phage display:**
  - $10^9$ – $10^{10}$ randomized protein sequences
  - Simultaneously tested for relevant biological function
Related Work: (Fromer et al. 2008)

Fromer M, Yanover, C. Bioinformatics (2008)
Related Work: (Fromer et al. 2008)

- Uses sum-product instead of max-product
- Obtain **per-position** amino acid **probabilities**
- Tried until convergence or 100000 iterations; all structures converged
Related Work: (Fromer et al. 2008)

• Conclusions:
  – Model results in probability distributions far from those observed experimentally.
  – Limitations of the model:
    • Imprecise energy function
    • Decomposition into pairwise energy terms
    • Assumption of a fixed backbone
    • Discretization of side chain conformations
for $m \leftarrow 1$ to $M$ do
  if $m = 1$ then
    $Cons^m \leftarrow \emptyset$
  else
    /* $t^m$, $p^m$, $q^m$ are the sub-space, position, rotamer to yield the next lowest energy sequence */
    $t^m \leftarrow \arg \max_{m' < m} \text{BMM}^{m'}$
    $a \leftarrow T(q^m)$  // aa type of $q^m$
  // Add pos. constraint to $Cons^m$:
  $Cons^m \leftarrow Cons^{t^m} \cup \{r^m \in \text{Rot}s^m | a\}$
  // Add neg. constraint to $Cons^m$:
  $Cons^m \leftarrow Cons^m \cup \{r^m \notin \text{Rot}s^m | a\}$
  Run BP to obtain: $\text{MM}_p(q) |_{Cons^m}$
  CalcBMM($t^m$)  // calculate BMM$^{t^m}$
end

Run BP to obtain: $\text{MM}_p(q) |_{Cons^m}$
for $i \leftarrow 1$ to $N$ do
  $r_i^m \leftarrow \arg \max_{r_i \in \text{Rot}s_i} \text{MM}_i(r_i) |_{Cons^m}$
  $S_i^m \leftarrow T(r_i^m)$  // $i^{th}$ aa of $m^{th}$ seq.
end
CalcBMM($m$)  // calculate BMM$^m$
return $\{S^m\}_{m=1}^M$