PREDICTING PROTEIN SIDECHAIN CONFORMATIONS – A NEW STRATEGY

IPiB Seminar - 30th March 2012
Sabareesh Subramaniam
Senes Lab, UW Biochemistry
PREDICTING TRANSMEMBRANE HELICAL DIMERS

ITLIIIFGVMAGVIGT
ITLIIIFGVRAGRIGT
SEARCH DIMER SPACE
TO EVALUATE EACH BACKBONE ORIENTATION - MODEL SIDECHAİNS
SIDECHAIN MODELING AS OPTIMIZATION

Backbone

Add sidechains to achieve minimum energy configuration
SIDECHAIN OPTIMIZATION

GLU  TRP  VAL
SIDECHAIN OPTIMIZATION
SIDECHAIN OPTIMIZATION

Energy = 40000 kcal / Mol
SIDECHAIN OPTIMIZATION

Energy = 30000 kcal / Mol
SIDECHAIN OPTIMIZATION

Energy = 70 kcal / Mol
SELECTED STRUCTURE – LOWEST ENERGY

Energy = 60 kcal / Mol
DISCRETIZED CONFORMATION LIBRARIES
DEGREES OF FREEDOM

Bond distances

Bond Angles

Dihedral or torsional angles

*Figures from Wikipedia
STATISTICS OF DIHEDRAL ANGLES

Side chain distribution

Relative density \((\times 10^{-3})\)

\(\chi_1\) vs. \(\chi_2\)
SIDECHAIN OPTIMIZATION

GLU

TRP

VAL
COMBINATORIAL SEARCH SPACE
(3-D JIGSAW PUZZLE)

No of conformations to search  36
COMBINATORIAL SEARCH SPACE
(3-D JIGSAW PUZZLE)

No of conformations to search  36* 54
COMBINATORIAL SEARCH SPACE
(3-D JIGSAW PUZZLE)

No of conformations to search  $36 \times 54 \times 3 = 5832$
Typically $> 10^{60}$
ALGORITHMS TO SEARCH ROTAMER SPACE

- Dead End Elimination
- Self Consistent Mean Field
- A* search
- Monte Carlo Simulated Annealing
- Graph decomposition
SO WHAT IS THE PROBLEM?
OUCH!
We need more conformations (sampling)
ROTAMER LIBRARY DETERMINES QUALITY OF SOLUTION

The template and the energy functions define a continuum energy landscape in side chain conformational space.

Ideal target
ROTAMER LIBRARY DETERMINES QUALITY OF SOLUTION

The template and the energy functions define a continuum energy landscape in side chain conformational space.

The rotamer library samples the landscape at discrete grid points.

Ideal target
ROTAMER LIBRARY DETERMINES QUALITY OF SOLUTION

The template and the energy functions define a continuum energy landscape in side chain conformational space.

The rotamer library samples the landscape at discrete grid points.

The search algorithm identifies the grid's minimum.

Ideal target

Best possible solution
We need more conformations (sampling)
MORE SAMPLING

Side chain distribution

relative density ($\times 10^{-3}$)
MORE SAMPLING

5x expanded library
GEOMETRIC FILTERS LEAD TO CONFORMER LIBRARIES

Conformers from high-res PDBs

Representative conformers
IGNORES THE NATURAL DISTRIBUTION

LEU

ASN
• People have been looking for solutions using the statistical distribution in structures

• However, the problem is that sampling is related to the energetics in a way that is difficult to predict
• People have been looking for solutions using the statistical distribution in structures

• However, the problem is that sampling is related to the energies in a way that is difficult to predict

• Solution: use energetics to identify the best sampling strategy for side chain optimization
BUILDING AN ENERGY BASED CONFORMER LIBRARY
GOALS

Can we create a library that can outperform existing libraries in terms of speed and/or Energies?

Can we create a flexible library where the conformers are in some useful order?
Can we sort the conformers instead of extracting a fixed-size subset?

5,000 conformers from high-res PDBs

50 representative conformers
Can we sort the conformers instead of extracting a fixed-size subset?

5,000 conformers from high-res PDBs

Energy based sorting

Sorted list of 5,000 conformers
• Use *energetics* to identify the best sampling strategy for side chain optimization

Conformers
• Use **energetics** to identify the best sampling strategy for side chain optimization

Conformers
• Use **energetics** to identify the best sampling strategy for side chain optimization
Use **energetics** to identify the best sampling strategy for side chain optimization.
• Use **energetics** to identify the best sampling strategy for side chain optimization
• Use **energetics** to identify the best sampling strategy for side chain optimization
- Use **energetics** to identify the best sampling strategy for side chain optimization

<table>
<thead>
<tr>
<th>Environments</th>
<th>Conformers</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Conformer 1" /></td>
<td><img src="image2.png" alt="Conformer 1" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Conformer 2" /></td>
<td><img src="image4.png" alt="Conformer 2" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Conformer 3" /></td>
<td><img src="image6.png" alt="Conformer 3" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Conformer 4" /></td>
<td><img src="image8.png" alt="Conformer 4" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Conformer 5" /></td>
<td><img src="image10.png" alt="Conformer 5" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Conformer 6" /></td>
<td><img src="image12.png" alt="Conformer 6" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Conformer 7" /></td>
<td><img src="image14.png" alt="Conformer 7" /></td>
</tr>
<tr>
<td><img src="image15.png" alt="Conformer 8" /></td>
<td><img src="image16.png" alt="Conformer 8" /></td>
</tr>
<tr>
<td><img src="image17.png" alt="Conformer 9" /></td>
<td><img src="image18.png" alt="Conformer 9" /></td>
</tr>
<tr>
<td><img src="image19.png" alt="Conformer 10" /></td>
<td><img src="image20.png" alt="Conformer 10" /></td>
</tr>
<tr>
<td><img src="image21.png" alt="Conformer 11" /></td>
<td><img src="image22.png" alt="Conformer 11" /></td>
</tr>
<tr>
<td><img src="image23.png" alt="Conformer 12" /></td>
<td><img src="image24.png" alt="Conformer 12" /></td>
</tr>
<tr>
<td><img src="image25.png" alt="Conformer 13" /></td>
<td><img src="image26.png" alt="Conformer 13" /></td>
</tr>
<tr>
<td><img src="image27.png" alt="Conformer 14" /></td>
<td><img src="image28.png" alt="Conformer 14" /></td>
</tr>
<tr>
<td><img src="image29.png" alt="Conformer 15" /></td>
<td><img src="image30.png" alt="Conformer 15" /></td>
</tr>
<tr>
<td><img src="image31.png" alt="Conformer 16" /></td>
<td><img src="image32.png" alt="Conformer 16" /></td>
</tr>
<tr>
<td><img src="image33.png" alt="Conformer 17" /></td>
<td><img src="image34.png" alt="Conformer 17" /></td>
</tr>
<tr>
<td><img src="image35.png" alt="Conformer 18" /></td>
<td><img src="image36.png" alt="Conformer 18" /></td>
</tr>
<tr>
<td><img src="image37.png" alt="Conformer 19" /></td>
<td><img src="image38.png" alt="Conformer 19" /></td>
</tr>
<tr>
<td><img src="image39.png" alt="Conformer 20" /></td>
<td><img src="image40.png" alt="Conformer 20" /></td>
</tr>
</tbody>
</table>
• Use **energetics** to identify the best sampling strategy for side chain optimization
THE first conformer
THE second conformer

Conformers

Environments

1 - 0 3 1 1 0 2

1st

2nd
CONFORMERS IN PROPORTION TO DISTRIBUTION

5x expanded library

Energy-Based Library
A walk in Trp space
560 complete protein repacks.
The lower the energy, the better.
DIHEDRAL RECOVERY
DIHEDRAL RECOVERY RESULTS
ADVANTAGES OF ENERGY BASED LIBRARY

• Helps achieve lower energy structures

• Better dihedral recovery

• Unprecedented flexibility – the first 'n' conformers is probably the best set of 'n' conformers
How to balance sampling across residue types?
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Single Repack Coverage

LEU  
ASN  
TRP

Number of Conformers
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Single Repack Coverage

Number of Conformers
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Single Repack Coverage

Number of Conformers

LEU
ASN
TRP
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Single Repack Coverage

LEU  ASN  TRP

Number of Conformers
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Number of Conformers vs. Single Repack Coverage

- LEU (green)
- ASN (red)
- TRP (blue)

Graph shows the coverage of single repack conformers for different amino acid types as the number of conformers increases.
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Single Repack Coverage

Number of Conformers

LEU
ASN
TRP
NUMBER OF CONFORMERS FOR EACH AMINO ACID TYPE

Number of conformers at different levels

ARG  ASN  ASP  CYS  GLN  GLU
SEARCH DIMER SPACE
USE DIFFERENT LEVELS TO IMPROVE SPEED AND ACCURACY
RESULTS OF GLYCOPHORIN A
PERFORMANCE

- Predict a structure in 10 mins
- On a cluster with 128 cores
- Sidechain optimization ~ 3-8 seconds
- Glycophorin A
  - < 0.7 Å RMSD from NMR structure
RESULTS ON THE HUMAN GENOME

- High throughput screening of the human genome
  - **CSF2R** - NLGSVYIYVLLIVGTLVCGIVLGF
  - **KCNE1** - ALYVMLGFFFGLFTLGIMLSYI
  - **LECT1** - VVLISGAVLLLFGAIGAFYFW
  - **MEP1A** - QVHGSVLMVIGGTAGVIFLTFSIAIL
  - **NRP1** - ILITIIAMSALGVLLGAVCGVVL
CSF2R - NLGSVYIYVLLIVGTLVC\textcolor{red}{GIVLGFLF}
LECT1 - VVLISGAVLLLLFGAIGAFYFW
MEP1A - QVHGSVLMVIGGTAGVIFLTFSIIAIL
NRP1 - ILITIIAMSALGVLLGAVCGGVVL
Thank You

Alessandro Senes

Ambalika Khadria
Loren LaPointe
Ben Mueller

Center for High Throughput Computing
University of Wisconsin-Madison