Molecular Modeling 2018 -- Lecture 8

- Local structure
- Database search
- Multiple alignment
- Automated homology modeling
What is local structure?

Early in the process of folding (nsec timescale) local structures form in the polypeptide chain which guide the formation of tertiary structure.
beta turns

4-residues

Residue 1 hydrogen bonds to residue 4

Type I (most common). Oxygen points away, viewed clockwise.

Type II (less common). Oxygen points toward, viewed clockwise.
Backbone angles and preferred sequence of beta turns

Backbone angles ±30°

<table>
<thead>
<tr>
<th>Type</th>
<th>$\phi_{i+1}$</th>
<th>$\psi_{i+1}$</th>
<th>$\phi_{i+2}$</th>
<th>$\psi_{i+2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>-60</td>
<td>-30</td>
<td>-90</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>-60</td>
<td>120</td>
<td>80</td>
<td>0</td>
</tr>
<tr>
<td>VIII</td>
<td>-60</td>
<td>-30</td>
<td>-120</td>
<td>120</td>
</tr>
<tr>
<td>I'</td>
<td>60</td>
<td>30</td>
<td>90</td>
<td>0</td>
</tr>
<tr>
<td>II'</td>
<td>60</td>
<td>-120</td>
<td>-80</td>
<td>0</td>
</tr>
<tr>
<td>Vla1</td>
<td>-60</td>
<td>120</td>
<td>-90</td>
<td>0*</td>
</tr>
<tr>
<td>Vla2</td>
<td>-120</td>
<td>120</td>
<td>-60</td>
<td>0*</td>
</tr>
<tr>
<td>Vlb</td>
<td>-135</td>
<td>135</td>
<td>-75</td>
<td>160*</td>
</tr>
</tbody>
</table>

Glycine rules turn propensity

1. Position type \ position type \ 1 \ 2 \ 3 \ 4
2. I \ P \ D/N/S/T \ G
3. II \ P \ P \ G
4. VIII \ G/P \ P \ P
5. I' \ G \ G
6. II' \ G

*have cis-peptide bond at $i+2$

http://www.ebi.ac.uk
Other local structures: Helix caps

- Proline helix C-cap
- Glycine helix N-cap

Helix C-terminus

Gly termination of helix

Schellman motif

“Capping Box” (Ser-X-X-Glu)

http://www.cryst.bbk.ac.uk
Datamining for local structure motifs

Structures from non-homologous proteins (not same family) were data-mined for correlated sequence/structure patterns. Strongest correlations were called “folding initiation site” (I-sites) motifs.
"Diverging" type-2 turn. A 7-residue peptide forms this structure!  


Local structure motifs are marked by glycines and hydrophobic patterns

<table>
<thead>
<tr>
<th>Motif</th>
<th>Average boundaries</th>
<th>Pattern of conserved non-polar residues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mda (°) dme (Å)</td>
<td></td>
</tr>
<tr>
<td>Amphiathic α-helix</td>
<td>56 0.71</td>
<td>1-4-8, 1-5-8</td>
</tr>
<tr>
<td>Non-polar α-helix</td>
<td>54 0.58</td>
<td>1-4-8, 1-5-8</td>
</tr>
<tr>
<td>Schellman cap type 1</td>
<td>81 1.01</td>
<td>1-6-9-11</td>
</tr>
<tr>
<td>Schellman cap type 2</td>
<td>76 0.94</td>
<td>1-6-8-9</td>
</tr>
<tr>
<td>Proline α-helix C cap</td>
<td>92 1.07</td>
<td>1-2-5-8</td>
</tr>
<tr>
<td>Frayed α-helix</td>
<td>75 0.96</td>
<td>1-5-9-13</td>
</tr>
<tr>
<td>Helix N capping box</td>
<td>99 0.95</td>
<td>1-6-9-13</td>
</tr>
<tr>
<td>Amphipathic β-strand</td>
<td>89 0.87</td>
<td>1-3, 1-3-5</td>
</tr>
<tr>
<td>Hydrophobic β-strand</td>
<td>101 0.91</td>
<td>1-2-3</td>
</tr>
<tr>
<td>β-Bulge</td>
<td>100 0.97</td>
<td>1-4-6</td>
</tr>
<tr>
<td>Serine β-hairpin</td>
<td>94 0.76</td>
<td>1-8</td>
</tr>
<tr>
<td>Type-I hairpin</td>
<td>80 0.94</td>
<td>1-7-8</td>
</tr>
<tr>
<td>Diverging type-II turn</td>
<td>87 1.04</td>
<td>1-7-9</td>
</tr>
</tbody>
</table>

Local structure formation

- Short pieces of protein sample conformational space randomly, driven by the hydrophobic effect (mostly).
- Glycines provide points of greater flexibility.
Folding

Secondary Structure Elements (SSE): alpha helix or beta strand

Local

Secondary

Super-secondary

Tertiary

Quaternary

Initiation sites like beta-alpha-beta units, hairpins
If the sequence is similar, then the structure is similar.
Searching for a homolog of known structure.

Download "Sequence 1" from http://www.bioinfo.rpi.edu/bystrc/courses/biol4550/biol4550.html. Name it "strepto.fasta"

Open it in MOE

SEQ: Protein > Search > PDB

In the database search window, Load chain (select strepto)

Search

Choose the hit with the best e-value

Inspect the alignment

Load the alignment. Close the search window,

In SEQ window, color residues by similarity (bottom bar Residues > Similarity)
Sequence Database search

- Your sequence (**query**) is chopped up into 3-tuples.
- Every 3-tuple (**there are exactly 8000**) has its own look-up table, or **index**, of database locations (**pdbcode, chain ID, position**)
- **Hits** are chains with the most 3-tuples arranged along a diagonal on the **alignment matrix**, query vs hit.
- Hits are aligned to query using the **dynamic programming algorithm** (**Smith-Waterman**)
- The Dynamic Programming score is converted to a statistic, called the **e-value**.
e-value

- The number of times in a database search that you will get a random, non-homologous hit with the same score or better.
How do I know it's a good alignment?

- In NCBI Blast: Look for a low e-value (<< 1). Lower is better.
- Long strings of contiguous matches is good. Lots of indels, bad.
- Are large portions of the target sequence missing?
- Are the indels "one-sided"? (all deletions in one sequence, all insertions in the other)
How do I know it's a good alignment?

- Look at a multiple sequence alignment. In a good MSA, indel positions tend to be conserved.
- Look at positions around the indel. How conserved are they?
- Check the coverage. Is every part of the target aligned to a template?
After loading alignment.

bold sequence, has coordinates

not bold, does not have coordinates

**Residues>Similarity** colors well-aligned regions blue, poorly aligned red.

Poorly aligned region. Reddish in *Similarity* coloring. Gaps all over the place.
In what order do I model?

Modeling philosophy: MAXIMIZE USE OF THE INFORMATION IN THE TEMPLATE

This means...

- Assign coordinates based on homology first.
  - **Identity** first,
  - Then **similarity**.
- Then **deletions**.
- Then **insertions**.
- Then **extensions**.
### How do I model?

What actions do I take?

<table>
<thead>
<tr>
<th>Action</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity</td>
<td>keep coordinates. no change.</td>
</tr>
<tr>
<td>similarity</td>
<td>mutate sidechain</td>
</tr>
<tr>
<td>deletions</td>
<td>remove residues, make new peptide bond, energy minimize.</td>
</tr>
<tr>
<td>insertions</td>
<td>predict loop conformation, position a loop. Make two new peptide bonds, energy minimize.</td>
</tr>
<tr>
<td>extensions</td>
<td>predict extension conformation, position an extension. Make new peptide bond, energy minimize.</td>
</tr>
</tbody>
</table>
Exercise: Sequence editing practice

• Create a new sequence:
• SEQ: New / Text Editor
• Type ACDEFGHIKLMNPQRSTVWY
• Save: aa.seq (close Text Editor window)
• Open: aa.seq (type: raw_seq)
• Try “abacus” mode: middle-mouse
  • Make this alignment—>
• Use “side” move: shift-middle-mouse
  • Make this alignment—>
• Use “slide” move ctrl-middle-mouse
  • Make this alignment—>
Manual homology modeling

- Work on Homework 3