Evaluating Models

same/different versus right/wrong

confidence -- 2 types

Fixing errors
"Same/different" versus "right/wrong."

There are 2 dimensions to models: model vs template is something we can see. Model vs target is something we can’t see, but can only infer.

<table>
<thead>
<tr>
<th>Template vs model</th>
<th>Same</th>
<th>Different</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same</td>
<td>Conserved, probably functionally similar</td>
<td>Interesting functional differences</td>
</tr>
<tr>
<td>Wrong</td>
<td>Unnoticed changes. Overly conservative modeling.</td>
<td>Changes where there shouldn't be. Overzealous modeling.</td>
</tr>
</tbody>
</table>

Low RMSD: detailed differences

High RMSD: large-scale differences
Cartesian coordinate differences: RMSD

- RMSD = root mean square deviation

\[
\sqrt{\sum_{i=1}^{N} (\vec{x}_i - \vec{y}_i)^2} \div N
\]

By far, the most widely used and accepted metric for structural difference.

- Identity
- One bond length
- One residue
- Random RMSD depends on length
- ?
Internal coordinate differences complement Cartesian ones

- Internal coordinates = bond distances, bond angles, torsion angles
- Deviations indicate **local** functional differences.
- MDA = maximum deviation in backbone angles
- Protein segments with mda < 120° almost always have superimposable structures.
- Superimposable structures do not always have mda < 120°.
**Internal coordinate differences:**

**Distance Matrix Error**

- **DME = distance matrix error** (average or RMS)
  
  Distance matrix \( D^x_{ij} = \) distance from \( i \) to \( j \) in structure \( x \)

\[
\sum |D^x_{ij} - D^y_{ij}| \quad \text{or} \quad \sqrt{\sum (D^x_{ij} - D^y_{ij})^2}
\]

\[
\sum |D^x_{ij} - D^y_{ij}| = \frac{N(N-1)}{2}
\]

\[
\sqrt{\sum (D^x_{ij} - D^y_{ij})^2} = \sqrt{\frac{N(N-1)}{2}}
\]

“N choose 2” = the number of pairs possible with \( N \) items = \( N(N-1)/2 \)
As for any difference metric, we must have an alignment first. The alignment associates $D_{ij}^y$ with $D_{ij}^x$.

$D_{ij}$ may be measured from $C\alpha$ to $C\alpha$, or from $C\beta$ to $C\beta$. (In the latter case, if the residue is a Gly, then $C\alpha$ is used instead.)
Confidence should measure **correctness**

```
<table>
<thead>
<tr>
<th>Target vs model</th>
<th>Template vs model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right</td>
<td>Same</td>
</tr>
<tr>
<td>Wrong</td>
<td>Different</td>
</tr>
</tbody>
</table>
```

- **High confidence** when the target and template are the same and the model is correct.
- **Low confidence** when the target and template are different and the model is correct.
Confidence = the estimated probability of being right.

Physics-based confidence estimate:
Based on modeling experience, knowledge of stereochemistry, function, other factors, not statistics. Case specific.

Knowledge-based confidence estimate:
Based on statistics of known structures and repeated modeling experiments. Empirical, not theoretical. Not specific to one case.
Knowledge-based statistics: Ramachandran allowed regions

- Check for other amino acids outside the allowed regions.
- If it is an outlier, is it conserved? Then it's real.

Remedies for suspicious outliers:
1. energy minimize with restraint
2. Ignore it. Outliers happen. But watch out. Too many outliers makes the whole model suspect...

Courtesy of Jane & David Richardson
kinemage.biochem.duke.edu
Ramachandran plot: outliers should be rare
Knowledge-based confidence: positive phi angle at Glycine

- Glycines, lacking a C–beta, have a greater allowed Ramachandran region, including the "$\alpha_L$", or positive phi, region.
- 2-fold symmetrized statistics for Glycine $\phi\psi$ angles show a more realistic picture of the energy landscape.

Bet on $\alpha_L$  
\[ \text{XXXXX} \text{G} \text{XXXXX} \text{G} \]  
\[ \text{XXXXX} \text{G} \text{XXXXX} \text{N} \]  
\[ \text{XXXXX} \text{G} \text{XXXXX} \text{N} \]  
\[ \text{XXXXX} \text{G} \text{XXXXX} \text{D} \]  
\[ \text{XXXXX} \text{G} \text{XXXXX} \text{G} \]  

Bets are off  

Courtesy of Jane & David Richardson
Knowledge-based confidence: Proline phi angle

- Check for impossible phi angles at Proline positions.
  If you find one, there are two possible remedies
  (1) energy minimize it away
  (2) re-align the Pro.

never leave it like that.

Courtesy of Jane & David Richardson
Knowledge-based confidence: cis peptide bond at X–Pro

• “cis peptides” : \( \omega \) (omega) torsion angle may only be 180° or 0° (because of double-bond character), but 0° is highly disfavored (and therefore rare!) unless the residue following the peptide bond is a Proline. Why is this true?

• X = the residue before Pro. X = big (F,Y,W) favors the trans state.

---

\( \omega \) angle 180°

---

\( \omega \) angle 0°
**Knowledge-based statistics: Preferred rotamers**

• **Rotamers** are preferred sidechain conformations, found by clustering database sidechains. • **Rotamer** sets (libraries) may be coarse grained or fine grained (pull-down menu in Rotamer explorer). • **Rotamers** have intrinsic energies, due to local interactions.

![General rotamer library](image1)

![Specific rotamer library](image2)

![Lysine sideview](image3)

![Phenylalanine](image4)

![Serine](image5)

**Compute | Biopolymer | Rotamer explorer**

Allows modeler to test rotamer swaps.

**Compute | Biopolymer | Protein geometry, rotamer**

Finds side chains that need help.
Physics-based confidence: void regions

- Nature abhors a void.

Remedies:
(1) re-pack sidechains with rotamer explorer.
(2) add waters.
(3) energy minimize with distance restraints
(4) Leave it alone. Voids may be functionally important. See (Paredes et al, BMC Bioinformatics 2011)
Physics-based confidence: buried charges

- Charges hate to be de-solvated.

Remedies:
(1) re-pack sidechains. Find a salt bridge.
(2) re-align. Put it on the outside.
(3) Leave it alone.

water dipoles delocalize the charge

buried charge is like a charge in a vacuum.
Build a homology model for a target (1BBZ.C) that actually has a known structure** using a different structure as template (2IIM.A)

(Why model it, since the structure is already known? you ask. Answer: To evaluate your modeling ability of course.)

**Please don't cheat by using the true structure as the template. (I will know if you do!)
Download the file linked to homework5 from the website (homework5.moe).

1BBZ.C will be the target (we will use only the sequence), and 2IIM.A will be the template.

2) Align the sequences/structures using the defaults (BLOSUM62, gap start 10).

3) Locate the two loops where there are indels. (One is near "NHNG", the other near "TKN"). Find the best anchor points and unalign all of the residues between the anchor points.


5) Open the file you just saved. It is chain #3. Drag it to position #1. It is now chain #1. Rename it "target".

6) Align chain #1 "target" to identical chain #2 without moving Chain #2. Then delete Chain #2 "1BBZ".

7) Save the MOE file. Upload it as Homework cb1 first part


Name the mdb file “hw2.mdb”
Selecting a loop

(9) Close all. Open homework5.moe. Delete 2IIM.
    (Keep 1BBZ, rename its tag “true target”.)
(10) Open the database file “hw2.mdb”. Opens in database viewer (DBV).
(11) In DBV, File | Browse
(12) Ribbon | (line-trace) ∨∨ , chain
(13) Hide | All atoms
(14) Rotate so as to view the NHNG loop (see figure) Browse.
(15) Find the loop that best matches the true target.
(16) Click Keep in the Browse window.
(17) Close Browse window.
(18) Inspect starting from the anchors. If you see beta-sheet H-bonds that are almost made, make them by adding distance restraints. This extends the β sheet.

(19) Near the middle of the hairpin must be a β-turn. **Type I or Type 2?** Look at the sequence (NHNGE) and the model.

<table>
<thead>
<tr>
<th>position type \ type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>P</td>
<td>D/N/S/T</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>P</td>
<td>P</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>III*</td>
<td>G/P</td>
<td>P</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>I'</td>
<td>G</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>II'</td>
<td>G</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(20) In Type-1 and Type-2 β turns, Gly should have a positive Φ angle. Does it? Measure by eye or check it using **Measure | Dihedral**

(21) If you think Gly should be in the 3rd position of a turn, make it Type-1 and make a H-bond restraint to enforce it. Or, if you think it is in the 4th position of a turn, then make it a Type-2 turn. Which is it?

(22) Restrain the Gly Φ angle to be in the range +60°,+120°
(23) Delete the true target. Keep the model. (Having both models would cause energy minimization to "explode"!!) Set force field to Amber12.
Compute | Prepare | Structure preparation
(24) Unfix only residues GYNHNGE. (Select. Unfix. Invert. Fix.)
(25) SVL: run ‘gizmin.svl’
Move atoms shift-meta-middlemouse, to help make the turn.
(26) Cancel | Minimize
(27) Open homework5.moe. Delete 2IIM. Align target and "true target" 1BBZ. Leave restraints in place. Don't delete them.
(28) Save and upload the MOE file as Homework cb2 second part