Protein Structure III
Protein Structure Prediction

- Protein **structure**...
  - ...is critical to protein **function**

- Problem:
  - protein structures are hard to determine
  - X-ray crystallography, etc., helps, but...
    - Only 10s of thousands of protein structures are known
    - Yet 100s of thousands of proteins are identified
Are protein **structures** determined by protein **sequences**?

- This would make a good debate...

- Or think-pair-share...

- Or at least a vote...
Does sequence determine structure? Some analogies

- Configuration of
  - collections of atoms
    - can determine structure

- Configuration plus environment of
  - collections of atoms
    - can determine structure

- Which holds for DNA? Protein? Water?
Does sequence determine structure? Some analogies

- **Configuration** of atoms can determine structure
  - **Examples:**
    - a water molecule
    - a fork
    - a hair, nail, hoof, feather, or horn of *keratin* protein

- **Configuration plus environment** of atoms can determine structure
Keratin: from configuration to structure

- Keratin protein makes hair, nails/claws, hooves, horns, beaks, feathers, turtle shell...
  - Not bones (collagen+calcium phosphate)
  - Not shells (calcium carbonate)
  - Not exoskeletons (chitin)
  - Dinosaur claws undoubtedly were keratin
    - Manning et al., Biological Letters 2006, 2(1):110-112
      - http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1617199/
  - Keratin is tough and hard, but a protein
    - Alpha and beta versions (based on helices and sheets)
  - What amino acid is especially high in keratin?
    - (lots of them in the keratin, or lots of keratin in them?)
Keratin (cont.)

- Keratin has many **disulphide** bonds
  - These lock into place the alpha helices and/or beta sheets
  - What amino acid is involved here?
- Human hair is about 14% that amino acid
Keratin (cont.)

- Human hair is about **14% cysteine**
- Cysteine is a food additive
  - Improves dough texture
  - It is often obtained from actual **human hair**, and added to e.g. **pizza crust dough**

*Bon appetit!*
Low-Cysteine Diet?

- Imagine a diet without cysteine...
  - What might happen?
Imagine a diet without cysteine...
- Not necessarily a problem!
  - It is a **non-essential** amino acid!
- What is the difference between:
  - Essential amino acids
  - Non-essential amino acids

But we’re not out of the woods
- A diet can be too low in sulfur generally
Sulfur Deficiency

- What might happen?
- Let’s all take a couple of minutes to find some fact about it on the Web
Does sequence determine structure? Some analogies

- Configuration of atoms can determine structure
  - Examples: (1) a fork; (2) a hair, nail, hoof, feather, or horn of keratin protein

- Atomic configuration plus environment can determine structure
  - Examples:
    - vacuum-detecting pop-up button on jar lids
      - Note the complex secondary structures involved
    - globular proteins (hydrophilic on outside)
    - what about integral membrane proteins?
    - what about fibrous proteins?
Does sequence determine structure? Some analogies

- **Configuration** of atoms can determine structure
  - Examples: (1) a fork; (2) a hair, nail, hoof, feather, or horn of keratin protein

- **Atomic configuration PLUS environment** can determine structure
  - Examples:
    - vacuum-detecting pop-up button on jar lids
      - Note the complex secondary structures involved
    - globular proteins (hydrophilic on outside)
Does sequence determine structure? More analogies

- Consider an example:
  - C-clamp plus playing card assembly
    - Note the complex secondary structures
    - How many stable configurations does it have?
Does sequence determine structure? More analogies

- c-clamp plus playing card assembly
  - How many stable configurations does it have?
- 2, or maybe infinity if you include turning the handle
  - if the handle is taped into place then it is two
- Can you think of a protein with 2 stable configurations?
Some proteins with 2 stable configurations

- Prion proteins
  - The most recently discovered broad category of "germ"
    - They’re not alive (so e.g. not bacteria)
    - They’re not viruses (though viruses not alive)
    - They are bistable proteins!
  - They cause certain exotic and deadly diseases...
Some proteins with 2 stable configurations

- Prion proteins
  - They are bistable proteins

- Prions cause exotic and deadly diseases...
  - The “wrong” fold spreads to neighboring molecules
  - Think of falling dominoes
    - After a while symptoms occur
      - Typically followed by death
Prions: bistable proteins

Prions can cause fearsome diseases

- Transmissible spongiform encephalopathies (TSEs):
  - Sporadic fatal insomnia (inherited form: fatal familial insomnia)
    - "Protein misfolding, not mutant gene, key to lethal sleep disorder" - www.uchospitals.edu/news/1999/19990526-sfi.html
  - Scrapie
  - Bovine spongiform encephalopathy (BSE)
    - Aka mad cow disease
  - Creutzfeldt-Jakob disease (CJD)
  - New variant Creutzfeldt-Jakob disease (CJDnv)
    - Basically CJD from eating beef
  - Kuru
    - Caught from eating dead people
  - Transmissible mink encephalopathy
  - Chronic wasting disease (deer, elk)
27 (why stop at 2?)

- http://www.technologyreview.com/biomedicine/24084/
  - (see hard copy also)
Structure from Sequence?

- The dream:
  - predicting structure from sequence
- So far, it is an impossible dream!
  - So other techniques are used instead
  - But maybe some day...
So how to do it, then?

Instead of “ab initio” prediction, use…

“knowledge based methods” (Westhead et al. p. 150)

Given: a target we want to find the structure of

Method 1: use **comparative modeling**
  - Compare it to proteins whose structure is known
  - Called comparative modeling

Method 2: use **secondary structure prediction**
  - Use knowledge about how amino acids provide evidence for the secondary structures they are in
Quick review:
- Is similar sequence evidence of similar structure?
- What evolves faster, generally speaking: sequence or structure?
Suppose we have a target and a template...

...and of 80 consecutive AAs, 25% match identically.

Then probably the structures are shared:

\[ t(\ell) = 290.15\ell^{-0.562} = 290.15 \times \frac{1}{80^{0.562}} = 24.72\% \]

- \( t(\ell) \) is the % of identically aligned amino acids required to conserve structure
- where \( \ell \) is the length of the sequence

This means we have found...

...the structure of the target and it is...
the same as the structure of the template.
Finding the Best Template

- What is a template?
- How might we find the best one?
Finding the Best Template...

...how to do it

- Use a database of sequences with known structures
- Find the best aligning template(s)
- That's your template!
  - If we meet the 80 AAs/25% criterion...
  - Then assume the structures are similar
Finding the Best Template II

- Problem:
  - Not enough templates are known
  - If there is no suitable template,
    - you pretty much can’t use comparative modeling
- “Solution”:
  - As time goes by,
    - more templates become known
    - (Or, use a different method for structure prediction)
Believability of the Prediction

- This depends in part on what % of AAs match
  - 70% and over – structures predicted well
  - Lower % - increased need to do manual reality checking
    - For example, specific AAs might be structurally important
      - Like what?
    - These should align even if the % decreases
How It’s Done

- If one template
  - Its structure is the hypothesized structure

- If multiple templates
  - Average the positions of corresponding AAs
  - That average is the hypothesized structure
    - Averaging can weight templates differently
    - Why might you weight one template more than another?
How It’s Done II

- Align the secondary structures
  - You know these for the template
  - The alignment with the target is guided
    - So the pure sequence alignment might differ
- Structure of loops:
  - Not as conserved
  - Harder to do
  - Use a “spare parts” database to find a similar loop from otherwise unrelated proteins
How It’s Done III

- Now predict side chain positions
  - Do this by finding a dense packing
    - Non-trivial algorithms can do this
- Finally, adjust the positions by minimizing the energy
  - Another non-trivial algorithm
Comparative modeling is good...
  - but suitable templates might not exist
  - Then, another approach is necessary

Secondary structure prediction is good...
  - does not use templates
  - It makes weaker predictions

Comparative modeling...
  - says where atoms are

Secondary structure prediction...
  - classifies subsequences as helical, strand, or coil
Predicting Secondary Protein Structure II

- Predicts amino acids’ secondary structure

- Meaning,
- the amino acid is in a secondary structure?
  or
- has its own secondary structure?
Predicting Secondary Protein Structure II

- Predicts secondary structure AA is in
  - classifies AA as helical, strand, or coil
  - $\alpha$-helices contain AAs in “helical” class
  - $\beta$-sheets:
    - made of strands
    - Contain AAs in “strand” class
  - coils contain AAs in “coil” class

- These classifications are often called...
  - three-state predictions
Secondary Structure Prediction

- We’ve looked at what it is
- Now let’s look at how it’s done
  - Amino acids have **propensities** to be helical or strand
  - Average propensity to be helical is 1.0
  - Average propensity to be strand is 1.0
- Are above average propensities **above** 1.0?
- Are below average propensities **above** 1.0?
Propensities of Amino Acids to Occur in Secondary Structures

Here is a way to compute propensities

- (see next slide if pdf messed up formula)
- Variable $A$ is an amino acid (# values?)
- Variable $S$ is a secondary structure
- $P$ is probability
- $|$ means “given”

Propensity formula

Here is a way to compute propensities:

Propensity \( A, S \) = \( \frac{P( A \mid S )}{P( A )} \)
The formula: Suppose a randomly selected amino acid in a helix is 1.59x more likely to be a Glu...than... an amino acid randomly selected from anywhere in a protein

What is Propensity(E, helix) ?

Here are the propensities....
Table 1. Helical and strand propensities of the amino acids.

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Helical (α) propensity</th>
<th>Strand (β) propensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLU</td>
<td>1.59</td>
<td>0.52</td>
</tr>
<tr>
<td>ALA</td>
<td>1.41</td>
<td>0.72</td>
</tr>
<tr>
<td>LEU</td>
<td>1.34</td>
<td>1.22</td>
</tr>
<tr>
<td>MET</td>
<td>1.30</td>
<td>1.14</td>
</tr>
<tr>
<td>GLN</td>
<td>1.27</td>
<td>0.98</td>
</tr>
<tr>
<td>LYS</td>
<td>1.23</td>
<td>0.69</td>
</tr>
<tr>
<td>ARG</td>
<td>1.21</td>
<td>0.84</td>
</tr>
<tr>
<td>HIS</td>
<td>1.05</td>
<td>0.80</td>
</tr>
<tr>
<td>VAL</td>
<td>0.90</td>
<td>1.87</td>
</tr>
<tr>
<td>ILE</td>
<td>1.09</td>
<td>1.67</td>
</tr>
<tr>
<td>TYR</td>
<td>0.74</td>
<td>1.45</td>
</tr>
<tr>
<td>CYS</td>
<td>0.66</td>
<td>1.40</td>
</tr>
<tr>
<td>TRP</td>
<td>1.02</td>
<td>1.35</td>
</tr>
<tr>
<td>PHE</td>
<td>1.16</td>
<td>1.33</td>
</tr>
<tr>
<td>THR</td>
<td>0.76</td>
<td>1.17</td>
</tr>
<tr>
<td>GLY</td>
<td>0.43</td>
<td>0.58</td>
</tr>
<tr>
<td>ASN</td>
<td>0.76</td>
<td>0.48</td>
</tr>
<tr>
<td>PRO</td>
<td>0.34</td>
<td>0.31</td>
</tr>
<tr>
<td>SER</td>
<td>0.57</td>
<td>0.96</td>
</tr>
<tr>
<td>ASP</td>
<td>0.99</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Why might an AA have both values <1?

(Caution: Westhead et al. give no formula so consistency between table and foregoing formula not guaranteed)
From Propensities to Structure

- Propensities of individual AAs are too weak to reliably predict structure.
- We need to combine evidence from a number of AAs to be reliable.

How might you propose doing this?
- (Let’s do a think-pair-share...)
From Propensities to Structure II

- **Chou-Fasman method**
  - 6 consecutive AA residues are in a *helix* if:
    - 4 are “helix favoring” (Westhead et al. p. 157),
    - their average helix propensity > 1, *and* their
    - ave. helix propensity > ave. strand propensity
  - AAs after the 6 are also in the helix until:
    - a proline occurs, *or*
    - 4 AAs in a row with helix propensity < 1 occur
Chou-Fasman method

5 consecutive AA residues are in a **strand** if:
- 3 are “strand favoring” (Westhead et al. p. 158),
- their average strand propensity > 1.04, and
- ave. strand propensity > ave. helix propensity

AAs after the 5 are also in the strand until:
- 4 AAs in a row with strand propensity < 1 occur
GOR (Garnier Osguthorpe Robson) method

- The structure that an AA is in depends on
  - its propensities
  - the propensities of its 8 preceding neighbors
  - the propensities of its 8 succeeding neighbors

- Review Q: Is the structure of an AA:
  - the larger structure it is in, or
  - the structure it possesses itself?
Accuracy of GOR and Chou-Fasman (early methods)

- Often <= 60% AAs classified correctly
  - Not exactly mind-blowing!
    ...compare...

- Suppose a protein was made of:
  1/3 helix, 1/3 strand, and 1/3 coil AAs
  - Predict their classes as follows:
    - throw a 3-sided die (i.e., guess wildly)
  - What % of AAs will be classified properly?
Improving the Accuracy

- 1st generation methods can be augmented with 2nd generation improvements
  - Basic idea: use MSAs to inform the process

- Given a MSA column with different AAs
  - If they share the same high propensity...
    - then that propensity is good evidence
    - Is propensity a physico-chemical property?
Some helices have hydrophobic and hydrophilic faces

- A helix can indeed have “faces”
- Then, hydrophilic/-phobic AAs form a repeating pattern
  - (the average length of a full turn is 3.6 AAs)
- If AA MSA substitutions preserve the pattern
  - ...that is evidence of a helix
Improving the Accuracy III

- In what secondary structure are insertions and deletions most common?
  - Use this fact to make secondary structure predictions

- Do you need a MSA to do this analysis?
2nd Generation Method Summary

- Use MSAs to provide more evidence
- A column that conserves propensity is significant evidence
- A repeating pattern of hydrophobic/-philic AAs that is conserved......significant
  - (period should be about 3.6)
- Insertions/deletions suggest coils
- These can raise accuracy to about 66%
The Most Modern Methods

- 66% isn’t enough...we need more!
- Modern methods use such methods as
  - Artificial intelligence
    - Machine learning
    - Artificial neural networks
- Accuracies over 70% are achievable
- Research continues
Predicting Integral Membrane Protein Structure

- Recall their general form:

What are the parts of this diagram??
Predicting Integral Membrane Protein Structure

- What secondary structures do integral membrane proteins tend to consist of?
- Membranes are lipid-based
  - So AAs that are within the membrane...
    - tend to be hydrophobic
    - long sequences of them suggest we have an integral membrane protein
      - Those are the transmembrane segments
  - How might we predict the connecting coils?
“Advanced” Structure Prediction

- *Ab initio* prediction
  - *ab initio* – Latin for “from the beginning”
  - The ‘t’ is pronounced either
    - ‘t’ (Latin pronunciation), or
    - ‘sh’ (anglicized pronunciation)
- This is fold prediction from first principles
  - Do you think it ought to be possible?
“Advanced” Structure Prediction

- *Ab initio* prediction
- (fold prediction from first principles)
  - Physics, chemistry, thermodynamics...
  - Do you think it ought to be possible?
Various people have worked on it
It sounds good
So far, it is not practically useful
Do you think it has the potential for the best performance?
How does nature determine protein structure?