Protein structure prediction

- Homology modeling (comparative modeling)
  - modeled based on homologous proteins
- Threading (fold recognition)
  - sequence – structure alignment
- de novo (free) modeling
  - No template is used

Roy & Zhang (2012)
**FUGUE (Version 2.0)**

**Profile Library Search Against HOMSTRAD**

- **Main Chain Conformation & Secondary Structure, Solvent Accessibility, Hydrogen Bonding Status**
- **Environment-Specific Substitution Table**
- **Structure-Dependent Gap Penalty**

*Gribble's profile (Lec 17)*

W(p, b) = \sum_{a,b} W(p, a) * Y(a, b)

- Profiles from structural alignments:
  - Environment-specific substitution table is used instead of BLOSUM62
- Higher gap penalties are given for the positions in secondary structures and core regions

**SwissPDBViewer**

- Coordinate data for a rough model can be viewed by SwissPDBViewer.

**Phyre2: Protein Homology/analogy Recognition Engine**

- protein threading
- HMM-profile alignments
- multiple sequence alignments
- clash interactions
- Ramsbottom distances between main chain atoms

### Protein structure prediction tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Method*</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-TASSER</td>
<td>TBM, FM</td>
</tr>
<tr>
<td>Robetta</td>
<td>FM</td>
</tr>
<tr>
<td>ModRef</td>
<td>TBRM</td>
</tr>
<tr>
<td>SWISS-MODEL</td>
<td>TBRM</td>
</tr>
<tr>
<td>HHPred</td>
<td>TBM, FM</td>
</tr>
<tr>
<td>Dali-TASSER</td>
<td>TBM, FM</td>
</tr>
<tr>
<td>QUARK</td>
<td>FM</td>
</tr>
<tr>
<td>Phyre</td>
<td>TBRM</td>
</tr>
<tr>
<td>SAM-TB</td>
<td>TBRM</td>
</tr>
<tr>
<td>TMHMM</td>
<td>TBM (meme-server)</td>
</tr>
<tr>
<td>LOMETS</td>
<td>TBM (meme-server)</td>
</tr>
<tr>
<td>PhyreR</td>
<td>TBM, TE</td>
</tr>
</tbody>
</table>

Protein structure prediction tools (Roy and Zhang, 2012; Dorn et al., 2017)

---

### NCBI: Structure

#### Show the VAST Structure Neighbors

**VAST: Vector Alignment Search Tool**

Structure Neighbors are determined by direct comparison of 3D protein structures.

http://toolkit.tuebingen.mpg.de/hhpred

http://bmbiist.tuebingen.mpg.de/ hhpred

MODeller output

http://bmbiist.tuebingen.mpg.de/ hhpred

Hildebrand et al. (2009) Proteins 75 (Suppl 9): 128-132

Gene Prediction: types of genes

- **Protein coding genes**
  - includes ORFs (open reading frames)

- **RNA genes** → do not code proteins
  - ribosomal RNA (rRNA)
  - transfer RNA (tRNA)
  - snRNA, etc.

Gene Prediction: protein coding genes

**[Prokaryotes]**
- High gene density → about 1 gene / kb
- Simple gene structure → no intron
- Short genes
- Overlapping genes

**[Eukaryotes]**
- Low gene density
- Complex gene structure → exons and introns
- Alternative splicing
- Pseudogenes

Open reading frame (ORF)

- A region with no stop codon (e.g., TAA, TAG, TGA)
  → can encode a part or all of a protein
- Starting by “Start” codon (e.g., Met)
  → not a requirement
- Secondary characters
  - Codon usage bias
  - Biased base composition
  → Periodicity of 3
  → Dicodon frequency

Open reading frame (ORF)

- No stop codon (e.g., TAA, TAG, TGA)

  → $P(\text{ORF})$: Probability of an ORF with $n$ codons

  \[
  P(\text{ORF}) = \left( \frac{61}{64} \right)^n
  \]

  (Probability of having no stop codon in $n$ codons based on universal codon table)

  \[
  P(20) = \left( \frac{61}{64} \right)^{20} = 0.38
  \]

  \[
  P(60) = \left( \frac{61}{64} \right)^{60} = 0.056
  \]

  \[
  P(100) = \left( \frac{61}{64} \right)^{100} = 0.008
  \]

  \[
  P(200) = \left( \frac{61}{64} \right)^{200} = 10^{-4}
  \]

  Not very likely to happen by chance
ORF Finder at NCBI

FramePlot
http://watson.nib.gov/junco/cg-bin/frameplot.pl

FramePlot
Distribution of start and stop codons
Plots of GC% at the 3rd codon positions from three different frames

FramePlot
Highly biased base composition (e.g., high GC%) in one frame relative to other frames may indicate protein coding gene regions!
**Base composition: coding exons vs. introns**

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Probable synonymous sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adh mRNA</td>
<td>31.4</td>
<td>19.4</td>
</tr>
<tr>
<td>Adh cDNA</td>
<td>31.4</td>
<td>19.4</td>
</tr>
<tr>
<td>Adh cDNA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In *Drosophila* genomes, synonymous sites (3rd codon positions) are often highly GC rich, but introns are AT rich.

**FramePlot**

- **D. melanogaster Adh cDNA**
  - Very high GC% in one frame
  - Other frames do not show the same bias.

**Base composition of *D. melanogaster* introns**

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syn</td>
<td>31.4</td>
<td>19.4</td>
<td>30.1</td>
<td>18.9</td>
</tr>
</tbody>
</table>


For *D. melanogaster*, GC% for Adh (81%) is significantly higher than for Adhr (59%). In all frames, synonymous sites are more GC-rich than non-coding regions.

**FramePlot**

- Introns & Non-coding:
  - High and similar AT% in all frames

- Exons: long gap between stop codons & high GC% in one frame
